Bias-Reduced Doubly Robust Estimation

Karel Vermeulen

Proefschrift voorgedragen tot het behalen van de graad van
Doctor in de Statistische Data-Analyse
Academiejaar 2014-2015

Promotor:
Prof. Dr. Stijn Vansteelandt

Vakgroep Toegepaste Wiskunde, Informatica en Statistiek
Faculteit Wetenschappen, Universiteit Gent
Krijgslaan 281, B-9000 Gent

"It takes no compromise to give people their rights... it takes no money to respect the individual. It takes no political deal to give people freedom. It takes no survey to remove repression."

– Harvey Milk

Table of Contents

Dankwoord

Het is nu bijna vier jaar geleden dat ik mijn diploma van Master in de Toegepaste Wiskunde in ontvangst mocht nemen. Op zo een moment denk je, *"nu weet ik toch al heel wat"*. Dat was echter een beetje naïef. Zoveel jaren later weet ik wel beter. De voorbije vier jaar heb ik dan ook enorm veel bijgeleerd. Het resultaat van dit leerproces is deze thesis. Heel wat mensen hebben me zowel op professioneel als persoonlijk vlak bijgestaan en op de een of andere manier mee bijgedragen tot het schrijven van deze thesis. Hierbij wil ik deze personen dan ook heel graag bedanken.

In de eerste plaats wil ik Stijn bedanken. Zonder jou was niets van dit alles mogelijk geweest. Ik was 19 toen ik voor het eerst in je les zat, lang geleden (we worden een jaartje ouder...). Toen besefte ik nog niet wat voor belangrijke rol je zou gaan spelen in mijn leven. Je was zowel mijn promotor voor mijn bachelorproef als voor mijn masterproef, dus waarom ook niet voor een derde keer, maar dit keer voor mijn doctoraat. Ik ben je dan ook enorm dankbaar dat je het opnieuw zag zitten om mijn promotor te zijn. Ik wil je bedanken voor alle energie en tijd die je in me hebt gestopt en voor wat ik allemaal door jou heb bijgeleerd. Ik weet dat ik niet altijd de meest gemakkelijke student was (en ben) en dat ik vaak koppig kon zijn en je advies niet altijd ten harte wou nemen, maar toch bleef je altijd geduld hebben. Je leerde me zelfstandig werken, maar ik was bovendien ook altijd welkom op je bureau om vragen te stellen. Als je me kon helpen, dan deed je dat, ook al had je zelf nog bergen werk. Het voelde dan ook goed dat er iemand achter me stond, een promotor die veel vertrouwen in me had. Je gaf mijn zelfvertrouwen dan ook

telkens een nieuwe boost wanneer het even weer zoek was. Op professioneel vlak gaf je me veel kansen. Ik kon heel wat nieuwe plaatsen op deze wereld ontdekken die ik anders niet snel te zien zou krijgen en ik kon zo ook heel wat nieuwe en interessante mensen leren kennen. Zo leerde ik ook dat je best een grote naam bent binnen het vakgebied van de causaliteit. Ik ben duidelijk niet de enige die naar je opkijkt! Verder gaf je me ook de kans om mijn onderwijsvaardigheden verder te ontwikkelen, wat ik met heel veel plezier deed. Het was een hele leuke ervaring om zowel wiskundigen proberen te overtuigen hoe interessant statistiek kan zijn en om niet-zo-wiskundigen de relevantie van wiskunde binnen de statistiek te doen inzien. Tijdens mijn doctoraat leerde ik je niet alleen beter kennen op professioneel vlak maar ook als mens. Dit zorgde ervoor dat ik alleen maar meer naar je begon op te kijken. Ik kan het nog altijd niet goed geloven hoe je alles combineert en gedaan krijgt. Ik bewonder dan ook je levensvisie en ben blij dat ik met iemand kon samenwerken die een heel tolerante geest heeft en vrijheid belangrijk vindt. Bedankt, want dat had je soms wel nodig wanneer ik weer met een of ander verhaal op de proppen kwam. Ik heb geluk gehad dat ik met je kon samenwerken! En hopelijk is dit niet het einde.

Naast Stijn kon ik ook op heel wat andere collega's op de S9 rekenen. Eerst en vooral wil ik de andere leden van de onderzoeksgroep statistiek bedanken voor zowel onze informele seminaries als ook onze statistiek-etentjes. Het is leuk om in een groep te werken waar we niet enkel alleen achter ons bureau zitten maar ook op regelmatige basis samen aan tafel zitten. In het bijzonder wil ik heel graag Machteld bedanken, veel meer dan zomaar een collega. We legden de voorbije jaren een gelijkaardig parcours af, eerst wiskunde, dan statistiek. Ik ben blij dat we samen op heel wat conferenties waren, met Boston en New York als toppers. Jouw aanwezigheid maakte het allemaal wat aangenamer en vertrouwder. In de voorbije jaren kon ik niet enkel in je bureau terecht voor R-vragen, maar ook om nu en dan eens mijn hart te luchten en om, in al mijn enthousiasme, een minder-werk-gerelateerd verhaaltje te vertellen. Bedankt. Ook wil ik Bart VR bedanken, altijd open voor een mopje met een goede portie sarcasme. Voor mij was je toch altijd een beetje de mature onder de niet-proffen wiens advies ik graag ten harte nam. Johan en Sjouke maar ook Joke, ook bedankt voor de fijne momenten op conferenties. Koen, jij bent alleszins de hipste statisticus die ik ken!

Vervolgens wil ik ook de andere mensen van de TWIST-vakgroep bedanken. Zeker mijn allerbeste bureaugenoot (voor de eerste drie jaar dan toch), Nele. Ik vond het heel fijn om die met jou te delen. We hadden veel fijne momentjes samen op het bureau! Bovendien heel erg bedankt om me de kans te geven om samen met jou een hoofdstuk in een boek te schrijven. Als je nu Nele Verbiest en Karel Vermeulen samen intikt op Google, zal je altijd resultaten krijgen! Jammer genoeg was het zonder jou niet meer hetzelfde! Ik kon je alles vertellen zonder dat ik me verlegen hoefde te voelen. Ondermeer door jou aanwezigheid kwam ik met veel plezier werken. Gelukkig kwam je vervanger Bart VG! Bart, ik ben blij dat we onze groene en linkse passies konden delen met elkaar. Natuurlijk wil ik ook de andere bureaugenoten, die in de voorbije vier jaar mijn aanwezigheid moesten dulden, bedanken, Gustavo, Marjon, Mushthofa, Ludger en Sarah.

Catherine, jij verdient ook een groot woord van dank. De sociale drijfveer van onze vakgroep. Spelletjesavonden, TWIST-bbq, kerstmarkt, en niet te vergeten, alle TWIST-weekends, je deed het gewoon allemaal. Dank je voor je organisatorisch talent en om altijd als een zonnetje rond te lopen, altijd goedlachs! Jij bent een echte sfeer-maker, een must-have in de vakgroep. Charlotte, misschien komt het omdat je zelf een dochter hebt, maar voor mij had je altijd een moeder-achtig gevoel over je, een beetje veiligheid, bedankt. Dan wil ik ook mijn persoonlijk computer-genie bedanken. Herman, hoeveel keer heb je me geholpen met het zwart scherm en met het gebruik van de STEVIN supercomputer? Heel wat! Heel erg bedankt! En zoals elk computer-genie, moet ook jij soms stoom aflaten, en dat vond ik heel leuk onder de vorm van Super Mario op mijn Wii U. We zijn nog altijd niet op het einde!

Natuurlijk is er geen vakgroep zonder secretariaat. Ik wil jullie dan ook allemaal bedanken voor de administratieve en technische ondersteuning waar ik de voorbije jaren gretig gebruik van maakte. Ann, je was een fantastische meter. Nu en dan een verrassing voor Sinterklaas of voor mijn verjaardag, je bent de beste! Dankzij jou kreeg ik ook altijd netjes mijn geld terug als ik weer eens op conferentie was gegaan. Je stond bovendien altijd klaar voor een babbel, ondanks wat je zelf allemaal hebt moeten doorstaan. Hilde, ook jou ben ik heel dankbaar. Voor wat kon ik niet bij jou terecht? Ik ben de tel kwijt van hoeveel petten jij nu eigenlijk droeg. In mijn eerste weken op de vakgroep was jij het die me op mijn gemak deed voelen tussen al die nieuwe mensen. Je stond altijd klaar om te helpen. Gesprekjes met jou leidden dan

ook tot de meest absurde onderwerpen, zeker op vrijdagnamiddag. Super fijn.

Ik heb ook geleerd dat de vakgroep een komen en gaan is van doctoraatsstudenten. Soms wil je dan ook liever iemand wat langer houden. Gilles, Bert, Virginie, Jan, Stéphanie, Davy, Nele,... Dank je om deel geweest te zijn van de vakgroep. Gelukkig konden we elkaar nog heel wat terugzien op activiteiten die Catherine zo mooi organiseerde. Virginie, op onze tweewekelijkse badminton-wedstrijden was de spanning altijd te snijden! Bovendien was kerstboom-shoppen nog nooit zo leuk. Jij bracht dit letterlijk tot een nieuwe dimensie. En Bert, ik weet nu nog altijd niet wie je emotioneel rijk vindt.

Mama en papa, ik weet eigenlijk niet hoe ik jullie kan bedanken. Ik ben blij om jullie zoon te zijn. Ook al volg ik niet meteen de meest traditionele manier van leven, wijkt deze een beetje af van het conventionele en ben ik bovendien soms nogal eigenzinnig, ik ben blij dat jullie me altijd gesteund hebben en mij de voorbije jaren alle vrijheid hebben gegeven die ik nodig had. Mama, ik heb geen idee hoe vaak ik met je aan de telefoon heb gehangen om mijn hart te luchten. En toch een beetje sorry dat ik het altijd zo druk had en niet vaak genoeg naar Deerlijk afzakte. Papa, mede dankzij jou heb ik het nu zo goed en kan ik in mijn eigen huis wonen. Hiervoor heb ik geen woorden. Bovendien, jullie verwachtten niets van me terug. Ik kijk dan ook heel hard op naar je vastberadenheid en je doelgerichte toekomstvisie. Ik kan daar nog veel van leren. Annelies, ook jij hebt wel wat van mijn eigenzinnigheid moeten verdragen, zeker toen je hier nog in Gent woonde, maar het lukte precies! Dankjewel om zo een fijne tweelingzus te zijn. Je was er dan ook op bepaalde momenten wanneer ik het echt nodig had. Ik ben blij dat je aan de start van een fijn gezinnetje staat! Ik kijk dan ook uit naar alle verdere ontwikkelingen. Lukas, dankjewel voor alle lashulp die ik in de toekomst zal nodig hebben, want als er nu iets is dat ik echt niet kan... Bovendien moet ik stilaan ook beginnen beseffen dat mijn kleine broer eigenlijk mijn grote broer aan het worden is en dat ik waarschijnlijk in de toekomst nog heel wat van je zal kunnen leren. Jij hebt heel wat talenten waar ik alleen maar van kan dromen! Jij maakt me een trotste broer. Ik ben bovendien heel hard onder de indruk van je gitaar-talenten en je doorzettingsvermogen dat je hiervoor nodig hebt. Hierbij wil ik natuurlijk ook de rest van de familie bedanken. Moeke, Pekel, Meme, Helena, Sarah, Astrid, Yente en Aren, jullie ontbraken ook niet!

Doctoreren is niet alleen werken maar ook voldoende ontspannen. Daarom wil ik zeker de *Chatty Gaymen and Britney* (sorry Brecht!) bedanken. Die ontspannende sfeer was een grote hulp de voorbije tijd! Ik voel me vereerd om deel uit te maken van Britneys selecte groepje van Chatty Gaymen. Mijn lieverdje Brecht, jouw steun was nogal cruciaal de voorbije maanden. Ik weet dat ik vaak bezig was met schrijven en werken en dat ik na een lange dag wel eens een brombeer kon zijn, maar toch, jij gaf me een gevoel van rust en deed me er aan herinneren dat ontspannen belangrijk is en verplichtte me het werk te vergeten. We hebben al heel wat fijne momenten achter de rug en ik ben benieuwd welke er nog allemaal zullen komen. De liefde die je me geeft doet ongelooflijk veel deugd. Dankje! Je bent zo een warm persoon en doet me kippenvel en tranen in de ogen krijgen. Je bent super! Ik kijk heel hard uit naar al onze geplande tripjes in de nabije toekomst!

Dan wil ik ook het creatieve team achter deze cover bedanken! Dieter, bedankt om dit allemaal mooi in Photoshop en InDesign te fixen. Wree Veys heeft dat goed gedaan! Peam, bedankt om voor mij deze vette dubbel robuuste T-Rex te tekenen. Ik ben ongelooflijk onder de indruk van jouw tekentalenten.

Sara, ook jou wil ik bedanken. Mijn leven in Gent is onlosmakelijk ook aan jou verbonden. Ik wil jou niet alleen bedanken voor alle fijne momenten en tripjes de voorbije jaren, maar ook voor het nalezen van delen van deze thesis en de gedetailleerde feedback die je me dan gaf. Dit was zeer nuttig. Ik kan me inbeelden dat er leukere dingen zijn om in je vrije tijd te doen. Bedankt!

Dan wil ik ook nog het ganse Oxfam-team van de Lammerstraat bedanken! Een shiftje doen in de winkel was, en is nog altijd, een uitstekend middel om even tot rust te komen. Zeker als daar altijd een lekker kopje faire koffie bij kan worden gedronken.

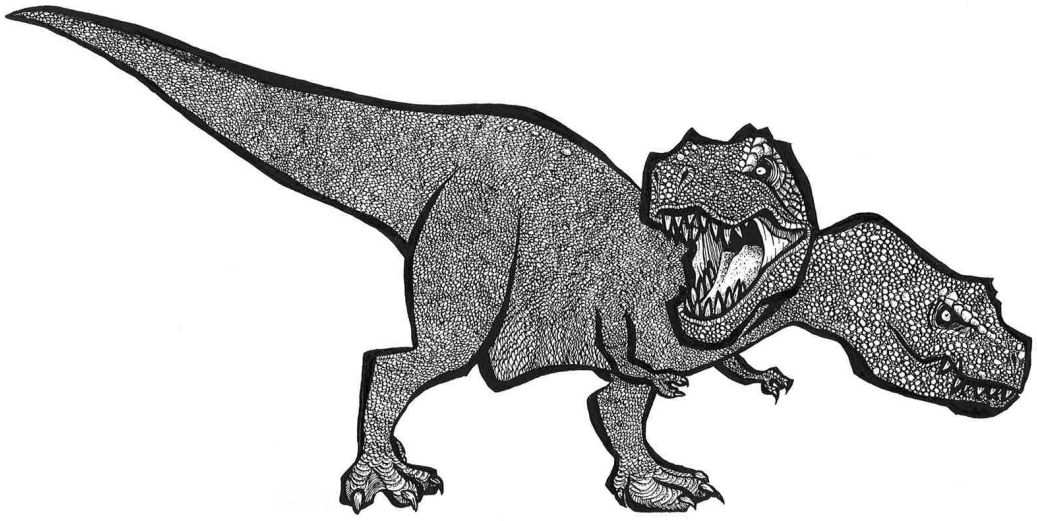
I would also like to take this opportunity to thank the members of the reading committee, Jan De Neve, Shaun Seaman and Arvid Sjölander. Thank you so much for reading my thesis in detail and thank you for the critical revision of this thesis and the interesting comments. I also would like to thank the other members of the examination committee, Els Goetghebeur, Tom Loeys, Olivier Thas and of course, Stijn Vansteelandt. Thank you all for being a member of the jury and for the time investment you made to read this work. Especially, I would like to thank Olivier for being the chair of the examination committee and for bearing

all the administrative burdens attached to this task. Additionally, I really enjoyed our fruitful collaboration concerning the permutation testing procedure for the extension of the Mann-Whitney test. Thank you.

Tenslotte wil ik het Fonds voor Wetenschappelijk Onderzoek (FWO) Vlaanderen bedanken voor het financieren van dit doctoraat en wil ik ook het STEVIN-team van de Universiteit Gent bedanken voor het gebruik van STEVIN supercomputer.

Ik kan het dan toch niet laten om te eindigen met een leerrijke uitspraak van mijn favoriete wereldverbeteraarster, Lady Gaga: *Don't you ever let a soul in the world tell you that you can't be exactly who you are. You were born this way.*

Karel Vermeulen
Mei 2015



List of Figures

List of Tables

CHAPTER 1

Introduction

Many empirical studies envision the assessment of the causal effect of a certain **treatment** or **exposure** A on some **outcome** of interest Y. Consider for instance an HIV trial where the aim is to study the efficacy of a new experimental drug (e.g., an augmented zidovudine therapy, $A = 1$) as compared to a standard treatment (e.g., zidovudine monotherapy, $A = 0$) on the CD4 cell count (cells/mm^3) at 20 weeks post-baseline (that is, measured 20 weeks after treatment initiation), see for instance Hammer et al. (1996). When μ_1 denotes the mean CD4 count (after 20 weeks) as if everybody in the study population received the experimental drug and μ_0 the mean CD4 count (after 20 weeks) as if everybody in the study population received the standard treatment, the target of inference is the difference $\mu_1 - \mu_0$. This difference

$$\mu_1 - \mu_0$$

indicates the **causal effect** of A on Y, measured on the additive scale. Randomized controlled trials are the gold standard for the evaluation of such cause-effect relationships. For this purpose, n (the sample size) HIV infected patients are randomized to either the experimental group ($A_i = 1$) or the group receiving standard treatment ($A_i = 0$). After 20 weeks, the intention is then to measure the outcome

Y_i for every patient $i = 1,\ldots,n$. When we succeed to collect these intended data $\{(A_i,Y_i) : i = 1,\ldots,n\}$ and correctly administer the randomization, we can unbiasedly estimate the causal effect $\mu_1 - \mu_0$ as

$$\sum_{i=1}^{n} A_i Y_i \Big/ \sum_{i=1}^{n} A_i - \sum_{i=1}^{n} (1-A_i) Y_i \Big/ \sum_{i=1}^{n} (1-A_i), \tag{1.1}$$

the difference of the group-specific averages of the outcomes. Unfortunately, practice teaches us that things are often not as easy as this and that many factors may prohibit analyses as easy as this one. For instance, humans are notorious for not doing what they are told to do. For example, it may happen that some study participants do not show up at scheduled clinical visits or do not completely fill in questionnaires. For these individuals, the outcome measurement may hence be missing and this often restricts the analysis to a selective subgroup of patients. This can lead to both a loss in precision and misleading conclusions. On the other hand, randomization is often not possible due to ethical and/or practical reasons. For instance, if interest lies in the effect of smoking behavior on the individual's survival time, it would be unethical to randomize the study participants to an exposure group that is forced to smoke or an exposure group for which smoking is prohibited. The lack of randomization makes the crude risk difference (1.1) potentially misleading because patients in both exposure groups may not be comparable: those that often smoke may also tend to consume more alcoholic beverages on average which may also have an effect on the survival time, confounding the relationship between the exposure (smoking behavior) and the outcome of interest (survival time). These problems are issues of **missing data** and **confounding**, which will be portrayed below in more detail. Furthermore, we will also see how the statistical community has developed advanced statistical methodology to remove bias due to selective missing data or confounding, but also where these methods might show some gaps, serving as a motivation for writing this thesis.

1.1 Doubly Robust Estimation in Missing Data Problems and Causal Inference

Below, we discuss in more detail why missing data and confounding are problematic and demonstrate how we can deal with these issues.

1.1.1 Missing data and doubly robust estimation

Recall the HIV example introduced before. Because the outcome is measured 20 weeks after treatment initiation, some patients may have dropped out and thus it is very likely that we will not be able to observe the outcome for all n patients. For these patients, the outcome measurement is **missing** (Schafer and Graham 2002; Carpenter and Kenward 2008). This can be formalized via the **missingness indicator** R_i, which equals one if we observe the outcome for the ith individual and equals zero otherwise. There can be various reasons why the outcome measurement is missing: some patients could have moved abroad or simply forgot their appointment at the hospital. However, patients can also miss their appointment because they were too ill to go to the hospital. Thus, in the presence of missing data, the observed data differ from the data we intended to collect; for this example, this can be written as $\{(R_i, A_i, R_i Y_i) : i = 1, \ldots, n\}$ and thus we only get to see the outcome Y_i whenever $R_i = 1$. To focus ideas, we restrict the attention to the estimation of μ_1. We furthermore suppose that the data are ordered in a way that the first $n_1 = \sum_{i=1}^{n} A_i$ patients in the sample are randomized to the experimental group (so $A_i = 1$ for $i = 1, \ldots, n_1$). A convenient choice to estimate μ_1 would be to use the **complete case** estimator

$$\sum_{i=1}^{n_1} R_i Y_i \Big/ \sum_{i=1}^{n_1} R_i \tag{1.2}$$

and thus averaging the outcome of those patients with both $A_i = 1$ and $R_i = 1$. This is different from the estimator that we would use if we would have collected all

intended data, which would equal

$$n_1^{-1} \sum_{i=1}^{n_1} Y_i = \sum_{i=1}^{n} A_i Y_i \Big/ \sum_{i=1}^{n} A_i. \tag{1.3}$$

Because we are forced to use only the data that we observe, we need to restrict the analysis to those patients with $R_i = 1$. There are now two important implications arising from these missing data. First of all, inevitably, there is a **loss of information** and consequently a loss in precision. The magnitude of this loss depends on the amount of missing data. This leads to larger standard errors of the estimators and potential lower power than intended for statistical hypothesis tests, e.g, to detect a causal effect. A second, and potentially more severe implication, is the risk of inducing **selection bias**: patients for whom we observe the outcome are not necessarily comparable to those patients for whom we do not observe the outcome. For instance, it may happen that patients with low CD4 count at baseline (and thus patients who are more critically ill at baseline) drop out during these 20 weeks because they feel too ill to participate in the study. In this case, the complete case estimator (1.2) will be systematically too large and overestimate the true mean μ_1 because we systematically get to see patients with larger CD4 counts. The complete case estimator is thus prone to selection bias and can lead to misleading conclusions.

To make progress, one has to assume certain assumptions about the reason why certain measurements are missing, the so-called **missingness mechanism**. One possibility would be that missingness is explainable by other auxiliary covariates $\boldsymbol{X}$, measured at baseline for all individuals. For instance, in the context of the HIV trial, besides typical information on each patient such as age, gender, weight, etc., we also might have collected a baseline CD4 count measure for everybody. A popular assumption is that these covariates $\boldsymbol{X}$ are sufficient to explain the reason of missingness. This assumption is what we refer to as the **missing at random** (MAR) assumption (Rubin 1976) and means that, given the observed data, missingness cannot depend on things we have not measured. In the current context, MAR specifically states that within subgroups of patients with a specific covariate pattern $\boldsymbol{X}$, the reason why some outcome values are missing is completely random and

thus that within such subgroups, the outcome Y is independent of the missingness indicator R:

$$Y \perp\!\!\!\perp R|\boldsymbol{X}. \tag{1.4}$$

This assumption is untestable (even with an infinite amount of data, one could not falsify this assumption); assessment of its validity and plausibility must rely on expert-knowledge.

Considering MAR to be plausible, generally speaking, we have two options to obtain a valid estimator of μ_1 (Robins et al. 1994): we can either exploit the outcome-covariate associations in a way to predict the missing outcome values, or alternatively, we can exploit the missingness-covariate associations, in a way to estimate the probability of (not) observing the outcome for a given covariate pattern. Nonparametric estimation of these associations is infeasible in the presence of a multi-dimensional covariate vector $\boldsymbol{X}$, which is known as the **curse of dimensionality** (Robins and Ritov 1997). We will therefore need to postulate so called **nuisance working models**, models that are not of scientific interest but needed to obtain a well-behaved estimator for the parameter of interest. We can thus either postulate a working model for the outcome-covariate associations such as a linear regression model, modeling the conditional mean of the outcome given the covariates among the responders (that is, those for which we observe the outcome); or, on the other hand, we can postulate a working model for the missingness mechanism such as a logistic regression model for the missingness indicator, modeling the probability of observing the outcome given the covariates. The first model could lead to a **regression imputation estimator**, which is based on averaging predictions for the outcome obtained via the working model for the outcome-covariate associations. In this case, the outcome measurements are imputed via working model predictions to obtain an estimator of μ_1. The second model could lead to a so-called **inverse probability of treatment weighted** (IPTW) estimator (Horvitz and Thompson 1952) which is based on inversely weighting each of the observed outcomes by the probability of being observed, given the covariates. This probability is calculated based on the working model for the missingness mechanism. Inverse weighting can be easily understood as follows. Suppose that for a certain covariate pattern, the probability of observing the outcome is 1/3. For this specific covariate pattern, we

thus expect to observe the outcome for one out of three such patients. By inversely weighting this patient's outcome by 1/3, this patient's outcome is counted three times and thus used as an outcome for two other patients with the same covariate pattern. As such, we construct a pseudo population of patients as if there were no missing data. These inversely weighted observed outcomes are then averaged to obtain an estimator for μ_1. Both of the aforementioned strategies rely on a single but different working model. By their reliance on a single working model, they only deliver a correct (**consistent** and **asymptotically unbiased**, that is, unbiased in large samples) estimator of the target parameter μ_1 when the corresponding working model is correctly specified.

A prevailing concern however is that in practice, misspecification of these nuisance working models will induce bias in the estimator of the target parameter μ_1 (Robins 1999a). This bias is called **model-misspecification bias** and is of a different nature than selection bias, discussed previously. The concern for bias due to model misspecification can be lessened via the use of so-called **doubly robust estimators** (Scharfstein et al. 1999a). These weaken the reliance on modeling assumptions by offering the opportunity to avoid committing to one specific modeling strategy in that they require specification of both of the aforementioned nuisance working models but remain consistent so long as one of both nuisance working models is correctly specified, regardless of which. They thus offer the data-analyst two chances for drawing valid inferences. Their reliance on both of these working models makes them potential compromise estimators amidst the regression imputation estimator and the IPTW estimator. Additionally, in the current context, a doubly robust estimator also makes efficient use of the available observed data in that it is **(locally) efficient** (Bickel et al. 1993a) within a broad class of estimators. It may therefore define the preferred analysis.

1.1.2 Doubly robust estimation in observational studies

We observed that missing data can seriously embroil the intended analysis but also how we can deal with this via the use of doubly robust estimators. We next describe another issue encountered in many empirical studies: the issue of **confounding** (d'Agostino 1998). Interestingly, this can be tackled in a very similar fashion as

the missing data issue. For this purpose, let us return to the initial set-up where we did measure the outcome Y_i for every individual but now consider the case where it was impossible to randomize patients to one of both treatment groups. This can be due to ethical and/or practical reasons. In this case, investigators are restricted to **observational studies**: they cannot manipulate the treatment mechanism but can only observe it. Estimation of the causal effect $\mu_1 - \mu_0$ is then often hindered by the possible presence of **confounders**: extraneous factors that are associated with both the exposure and the outcome, and thereby distort their association. For instance in the HIV study, suppose that the augmented zidovudine treatment is primarily given to patients that are most severely ill because it may be so that standard zidovudine treatment is known not to be helpful anymore for such patients and the augmented treatment might be the last hope for these patients. In this case, patients receiving the augmented treatment are not comparable to patients receiving standard treatment, making disease severity an important confounder. The estimator $\sum_{i=1}^n A_i Y_i / \sum_{i=1}^n A_i - \sum_{i=1}^n (1-A_i) Y_i / \sum_{i=1}^n (1-A_i)$, which merely measures a crude association between A and Y in an observational study, will then generally be badly biased for the true causal effect $\mu_1 - \mu_0$, so-called **confounding bias**, similar to selection bias. Nevertheless, the causal effect can still be estimated based on data from an observational study when we collect sufficient data on the confounders $\boldsymbol{X}$ for the outcome-exposure association. A causal effect can then be identified if one assumes that there are no unmeasured confounders, the **no-unmeasured confounders assumption** (Hernán 2004; Hernán and Robins 2006), in the sense that within a specific confounders stratum $\boldsymbol{X}$, it is as if treatment was randomly assigned. This assumption is thus similar to the MAR assumption. Indeed, for a patient receiving standard treatment, we observe its outcome under treatment condition $A = 0$. However, its outcome under treatment condition $A = 1$ is unobserved and can be considered as missing, so the problem of confounding can be translated to a missing data problem. The no-unmeasured confounders assumption is also untestable. Hence, its validity and plausibility must also be based on expert-knowledge.

In a similar fashion as for the missing data problem, one could either postulate a nuisance working model for the association between the outcome and confounders and exposure to obtain a regression imputation estimator for the causal effect or

one could postulate a nuisance working model for the probability of being treated, the **propensity score** (Rosenbaum and Rubin 1983), to obtain an IPTW estimator for the causal effect. Because these estimators are also based on a single working model, model misspecification bias also constitutes a severe concern. This concern can again be lessened via the use of a doubly robust estimator. Finally note that, on top of confounding, observational data can also be subject to missing data in for instance the outcome and one then has to combine the ideas put forward in this and the previous section.

1.1.3 Doubly robust estimation in randomized experiments

We noted that, besides enjoying the double robustness property, many doubly robust estimators also have desirable efficiency properties in the sense that they are **locally efficient** within a broad class of estimator. For instance, the doubly robust estimator of the risk difference $\mu_1 - \mu_0$ that adjusts for confounding in observational studies, is the most efficient estimator among the class of all estimators that are consistent when the propensity score working model is correctly specified, provided that also the working model for the association between the outcome and the confounders and exposure is correctly specified. The efficiency property is thus local. Because of this, the use of doubly robust estimator has also been advocated in randomized experiments (Tsiatis et al. 2008; Moore and van der Laan 2009; Vermeulen et al. 2015). This is because in this case, the propensity score (that is, the randomization probability) is known by design as it is under control of the investigator so that the doubly robust estimator is guaranteed to be consistent. Exploiting these known randomization probabilities makes it possible to increase the power of a statistical hypothesis tests aiming to detect a treatment effect via covariate adjustment without risking bias due to model misspecification.

1.2 Nuisance Working Model Estimation

Doubly robust estimators enjoy their defining property of double protection against model misspecification of the nuisance working models. Estimation of the **nuisance parameters** indexing these working models however has long been ignored. This

is because theoretical results show that the choice of nuisance parameter estimators has no impact on the asymptotic variance of the doubly robust estimator when both working models are correctly specified. This has lead to the default use of maximum likelihood estimation (MLE) of the nuisance parameters indexing the working models (Bang and Robins 2005).

However, despite their attractive properties, doubly robust estimators have been the subject of recent debate (Kang and Schafer 2007a; Ridgeway and McCaffrey 2007; Robins et al. 2007; Tan 2007; Tsiatis and Davidian 2007; Kang and Schafer 2007b). First of all, under misspecification of at least one working model, many doubly robust estimators for a given target parameter can be constructed by varying the choice of nuisance parameter estimators, all with potentially very different behavior under working model misspecification. This implies that more subtle choices for the nuisance parameter estimators can be made. Second, it is likely that model misspecification affects all working models in practice, and thus the very premise that at least one of both working models is correctly specified, lives on shaky grounds. Moreover, the performance of doubly robust estimators can sometimes be worse than that of competing estimators that do not enjoy the double protection property.

This has encouraged statisticians to identify nuisance parameter estimators which primarily aim at variance reduction under misspecificaton of one working model but also how to make clever use of data-adaptive learning algorithms. In this thesis, we will investigate the usefulness of doubly robust estimators from the perspective that all models are wrong. The fundamental objective is to develop a general estimation principle for the nuisance working models used in the construction of such doubly robust estimators (from the prospect where both working models are misspecified) where the focus is on **reducing bias**. This is motivated by the fact that the bias of a doubly robust estimator may become especially severe under misspecification of both working models (Kang and Schafer 2007a; Vansteelandt et al. 2012).

1.3 Organization of This Thesis

Because many of the results developed in this thesis heavily rely on the theory of **semiparametric models and semiparametric efficiency**, we briefly review the necessary semiparametric theory in **Chapter 2**, which is based on the geometry of influence functions in parametric and semiparametric statistical models. We furthermore apply this theory to the problem of estimating a population mean outcome with incomplete data, explainable by a set of measured auxiliary covariates. The resulting locally efficient, regular and asymptotically linear estimator of this target parameter will be used as an object of study throughout. The content of this introductory chapter is primarily based on the excellent book *Semiparametric Theory and Missing Data* by Tsiatis A.A. (2006), Springer: New York, which gives a more detailed overview of the semiparametric efficiency theory in missing data problems.

In **Chapter 3**, we demonstrate that the aforementioned locally efficient estimator also possesses a remarkable and attractive property: **double robustness**. This property states that the estimator consistently estimates the target parameter when either a working model for the missingness mechanism or a working model for the conditional mean outcome is correctly specified. Doubly robust estimation is not restricted to this missing data problem but is now also well established for many other statistical parameters. Their popularity can be judged from the many scientific articles on doubly robust estimation: over 2000 on Google Scholar; over 200 on Web of Science, in spite of the theory on double robustness being relatively new. Recently, they are also being considered by companies, such as Google and Microsoft, for policy optimization and evaluation in the context of content recommendation and internet advertising (Dudík et al. 2015).

However, in spite of their defining double protection property, Kang and Schafer (2007a) cautioned for potentially disastrous performance of certain doubly robust estimators (relative to simpler estimators) when one or both working models is/are misspecified. They moreover reveal that many different doubly robust estimators may exist for a given target parameter, all with potentially very different behavior and properties under misspecification of at least one working model. These concerns have encouraged statisticians to identify alternative nuisance parameter estimators,

which primarily aim at variance-reduction of the doubly robust estimator under misspecification of one working model but also how to make clever use of data-adaptive learning algorithms. In the first part of this thesis (in Chapter 4 and Chapter 5), we will also identify alternative nuisance parameter estimators for existing doubly robust estimators. However, we will primarily focus on **bias-reduction** rather than variance-reduction. In the second part of this thesis (in Chapter 6 and Chapter 7), we will focus on the construction of new doubly robust estimators and exploit their local efficiency property in the analysis of randomized experiments.

Specifically, in **Chapter 4**, we will investigate the usefulness of doubly robust estimators from the perspective that all models are wrong. This is motivated by the fact that the bias of a doubly robust estimator may become especially severe under misspecification of both working models. In particular, we will propose a simple and fairly generic estimation principle for finite-dimensional nuisance parameters indexing all working models with the defining property of locally minimizing the squared first-order asymptotic bias of the doubly robust estimator, referred to as **bias-reduced doubly robust estimation**. The bias-reduced doubly robust estimation principle introduced in Chapter 4 is confined to the use of parametric nuisance working models. To allow for further bias reduction, in **Chapter 5**, we will investigate how data-adaptive learning algorithms can be integrated in the biased-reduced doubly robust estimation procedure. The proposed procedure is referred to as **data-adaptive bias-reduced doubly robust estimation**.

As previously announced, in **Chapter 6**, we will exploit the local efficiency property, shared by many doubly robust estimators, in the context of the Mann-Whitney U test. Specifically, we will propose a **locally efficient estimator** of the **marginal probabilistic index** (MPI), the effect size considered by the classical Mann-Whitney U test. This will allow for flexible covariate adjustment so as to increase the power of the classical Mann-Whitney U test. However, because the Mann-Whitney U test is often indicated in small samples, where standard errors based on asymptotic approximations may not well approximate the true sampling variability, we also propose a **permutation test** based on this locally efficient estimator. Next, in **Chapter 7**, we extend the ideas put forward in Chapter 6 and present a **doubly robust adaptation of the Mann-Whitney U test** to adjust for confounding in observational studies. The resulting doubly robust estimator

yields a consistent estimator of the MPI under either correct specification of a working model for the propensity score or a working model for the conditional probabilistic index. We end this chapter with a discussion on different alternative nuisance parameter estimation strategies, constructed to enhance the performance of the novel doubly robust estimator. Specifically, we briefly outline how we could obtain a doubly robust regression imputation estimator, a bias-reduced doubly robust estimator (using the methods introduced in Chapter 4) and a doubly robust estimator based on empirical efficiency maximization.

We end this thesis in **Chapter 8** with a reflection on the results obtained throughout this thesis and a final conclusion. We furthermore make suggestions for future research.

CHAPTER 2

Semiparametric Theory

Many of the results developed in this thesis demand some understanding of the theory of semiparametric models and semiparametric efficiency. In particular, a formal understanding of the theory regarding the geometry of influence functions is needed. In this chapter, we will summarize the basic concepts. We start by introducing semiparametric models in Section 2.1 and some notions of Hilbert spaces are given in Section 2.2. Afterwards, in Section 2.3, we introduce the key features of the geometry of influence functions which will enable us to study the efficiency of estimators for finite-dimensional parameters describing parametric models. Subsequently, these concepts and results are extended to estimators for finite-dimensional parameters in semiparametric models in Section 2.4. We end this chapter in Section 2.5 by applying this theory to the estimation of a population mean outcome in the presence of incomplete data, explainable by auxiliary covariates. This particular problem will be studied in detail throughout this thesis and is formally equivalent to the estimation of a mean counterfactual outcome where a sufficient set of covariates is collected to adjust for confounding. For a more comprehensive overview of the literature concerning semiparametric models and semiparametric efficiency theory, we refer to Bickel et al. (1993b); Newey and

McFadden (1994); Tsiatis (2006).

2

2.1 Statistical Models

Statistical problems are formalized using probability models for the observed data. Consider a random sample of i.i.d. (independently and identically distributed) observations $\mathcal{O} = \{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$ of size n, with n the sample size. The vector $\boldsymbol{O}_i$ represents the observed data collected for the ith subject and can contain for example an outcome measure Y_i, an exposure A_i and a vector of covariates $\boldsymbol{X}_i$. The true (unknown) joint density function of a single $\boldsymbol{O}$ is denoted by $f_{\boldsymbol{O},0}(\boldsymbol{o})$. A statistical or probability model is then defined as a class of density functions of the observed data for which the researcher believes that might have generated the data (and thus includes $f_{\boldsymbol{O},0}$), indexed by finite and/or infinite-dimensional parameters. A **nonparametric model** is loosely speaking a statistical model which puts no restrictions on the class of densities of the observed data (except for some smoothness or moment conditions). This is denoted by

$$\mathcal{M}_{\text{NP}} = \left\{ f_{\boldsymbol{O}}(\boldsymbol{o}) \middle| f_{\boldsymbol{O}}(\boldsymbol{o}) \geq 0 \text{ for all } \boldsymbol{o} \text{ and } \int f_{\boldsymbol{O}}(\boldsymbol{o}) d\boldsymbol{o} = 1 \right\}. \tag{2.1}$$

To avoid technical difficulties, $d\boldsymbol{o}$ is used to denote an appropriate measure, which can be for instance the Lebesgue measure, the count measure or a combination of both.

When prior knowledge about the shape of the observed data distribution is available, the nonparametric model is too large. In this situation, the researcher might be willing to assume a **parametric model** $\mathcal{M}_{\text{P}} \subset \mathcal{M}_{\text{NP}}$, indexed by a finite-dimensional parameter $\boldsymbol{\theta}$:

$$\mathcal{M}_{\text{P}} = \{ f_{\boldsymbol{O}}(\boldsymbol{o}; \boldsymbol{\theta}) \in \mathcal{M}_{\text{NP}} | \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p \}, \tag{2.2}$$

with p the dimension of $\boldsymbol{\theta}$. Given that the model $\mathcal{M}_{\text{P}}$ holds, the truth $\boldsymbol{\theta}_0$ is such that $f_{\boldsymbol{O},0}(\boldsymbol{o}) = f_{\boldsymbol{O}}(\boldsymbol{o}; \boldsymbol{\theta}_0)$. Often, $\boldsymbol{\theta}$ can be partitioned as $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$ with $\boldsymbol{\beta}$ a q-dimensional parameter of interest and $\boldsymbol{\eta}$ an r-dimensional nuisance parameter

($p = q + r$). In other cases[1], no such partition is feasible and $\boldsymbol{\beta}$ can be seen as a function $\boldsymbol{\beta}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$.

Unfortunately, parametric models can be too restrictive and full specification of the joint density function is often not necessary. In this case, $\boldsymbol{\theta}$ will be infinite-dimensional but the statistical model will still put restrictions on the class of allowed joint density functions, in contrast to nonparametric models. We refer to this type of model as a **semiparametric model**:

$$\mathcal{M}_{\text{SP}} = \{f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\theta}) \in \mathcal{M}_{\text{NP}} | \boldsymbol{\theta} \in \Theta, \Theta \text{ an infinite-dimensional set}\}. \tag{2.3}$$

As for parametric models, $\boldsymbol{\theta}$ can often be partitioned as a q-dimensional parameter $\boldsymbol{\beta}$ of interest and an infinite-dimensional nuisance parameter $\boldsymbol{\eta}$. In other cases, no such partition is feasible and $\boldsymbol{\beta}$ can still be seen as a function $\boldsymbol{\beta}(\boldsymbol{\theta})$ of the infinite-dimensional $\boldsymbol{\theta}$. An example will be given in Section 2.5.

2.2 Hilbert Space for Random Vectors

A key feature in the development of semiparametric theory is the Hilbert space of mean-zero q-dimensional random functions, which is infinite-dimensional. The content of this section is primarily based on Tsiatis (2006, chap. 2).

A Hilbert space, denoted by $\mathcal{H}$, is a complete normed linear vector space equipped with an inner product.

> **Definition 2.1** (Inner Product)**.** *An **inner product** is a function $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that maps any two elements of $\mathcal{H}$ to a real number such that for any $\boldsymbol{h}_i \in \mathcal{H}$ $(i = 1,2,3)$ and $\lambda \in \mathbb{R}$, (1) $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = \langle \boldsymbol{h}_2, \boldsymbol{h}_1 \rangle$, (2) $\langle \boldsymbol{h}_1 + \boldsymbol{h}_2, \boldsymbol{h}_3 \rangle = \langle \boldsymbol{h}_1, \boldsymbol{h}_3 \rangle + \langle \boldsymbol{h}_2, \boldsymbol{h}_3 \rangle$, (3) $\langle \lambda \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = \lambda \langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle$ and (4) $\langle \boldsymbol{h}_1, \boldsymbol{h}_1 \rangle \geq 0$ with equality if and only if $\boldsymbol{h}_1 = \boldsymbol{0}$.*

The inner product allows us to define the norm, i.e., the length of vectors in $\mathcal{H}$, which is the distance to the origin.

[1] This representation will be particularly useful when we consider semiparametric models. For parametric models, we can always reparametrize the model so that there is a one-to-one relationship between $\{\boldsymbol{\beta}(\boldsymbol{\theta})^T, \boldsymbol{\eta}(\boldsymbol{\theta})^T\}$ and $\boldsymbol{\theta}$, for some r-dimensional nuisance function $\boldsymbol{\eta}(\boldsymbol{\theta})$.

Definition 2.2 (Norm)**.** *The **norm** based on an inner product $\langle \cdot, \cdot \rangle$ is the function $\| \cdot \| : \mathcal{H} \to \mathbb{R}^+$ defined as $\|\boldsymbol{h}\| = \langle \boldsymbol{h}, \boldsymbol{h} \rangle^{1/2}$ for any $\boldsymbol{h} \in \mathcal{H}$.*

The norm also serves as a distance function: the distance between two elements $\boldsymbol{h}_i \in \mathcal{H}$, $i = 1, 2$, is given by $d(\boldsymbol{h}_1, \boldsymbol{h}_2) = \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|$. The inner product also enables us to define the concept of orthogonality.

Definition 2.3 (Orthogonal Vectors)**.** *Two vectors $\boldsymbol{h}_i \in \mathcal{H}$ $(i = 1, 2)$ are said to be **orthogonal**, denoted $\boldsymbol{h}_1 \perp \boldsymbol{h}_2$, if and only if $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = 0$.*

Two other useful results, implied by the definition of the inner product, the related norm and orthogonality, are the following.

Theorem 2.1 (Pythagorean Theorem)**.** *For $\boldsymbol{h}_1 \perp \boldsymbol{h}_2$ with $\boldsymbol{h}_i \in \mathcal{H}$, we have that*

$$\|\boldsymbol{h}_1 + \boldsymbol{h}_2\|^2 = \|\boldsymbol{h}_1\|^2 + \|\boldsymbol{h}_2\|^2.$$

Theorem 2.2 (Cauchy-Schwartz Inequality)**.** *For any two elements $\boldsymbol{h}_i \in \mathcal{H}$, $i = 1, 2$,*

$$|\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle|^2 \leq \|\boldsymbol{h}_1\|^2 \|\boldsymbol{h}_2\|^2,$$

with equality if and only if $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are linearly dependent; that is $\boldsymbol{h}_1 = \lambda \boldsymbol{h}_2$ for some scalar $\lambda \neq 0$.

These definitions are now sufficient to define the Hilbert space of mean-zero q-dimensional random functions, which will also be denoted by $\mathcal{H}$. Consider a probability space $(\Omega, \mathcal{F}, P)$, with Ω the sample space, $\mathcal{F}$ the corresponding σ-algebra, and P the probability measure over the measurable space $(\Omega, \mathcal{F})$ that generates the observed data $\boldsymbol{O}$.

Definition 2.4 (Space of Mean-Zero q-dimensional Random Functions)**.** *Consider a q-dimensional measurable random function $\boldsymbol{h}(\boldsymbol{O}) : \Omega \to \mathbb{R}^q | \omega \mapsto \boldsymbol{h}(\boldsymbol{O})(\omega) = \boldsymbol{h}\{\boldsymbol{O}(\omega)\}$. Define the space $\mathcal{H}$ as the space of all such q-dimensional functions $\boldsymbol{h}(\boldsymbol{O})$ that have mean zero and finite second moment:*

$$\mathcal{H} = \left\{ \boldsymbol{h}(\boldsymbol{O}) \middle| E\{\boldsymbol{h}(\boldsymbol{O})\} = \mathbf{0} \text{ and } E\{\boldsymbol{h}^T(\boldsymbol{O})\boldsymbol{h}(\boldsymbol{O})\} < \infty \right\}. \tag{2.4}$$

The expectation is defined with respect to the true underlying probability measure P. When we refer to an element $\boldsymbol{h} \in \mathcal{H}$, we implicitly refer to the random function $\boldsymbol{h}(\boldsymbol{O})$.

Definition 2.5 (Covariance Inner Product)**.** *The **covariance inner product** for arbitrary elements $\boldsymbol{h}_i \in \mathcal{H}$, $i = 1,2$, is defined as*

$$\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = E(\boldsymbol{h}_1^T \boldsymbol{h}_2) = E\{\boldsymbol{h}_1^T(\boldsymbol{O})\boldsymbol{h}_2(\boldsymbol{O})\}. \tag{2.5}$$

Clearly, the covariance inner product satisfies the first three conditions of an inner product. As for the fourth condition, we define $\boldsymbol{h}_1$ to be equivalent to $\boldsymbol{h}_2$ iff $P(\boldsymbol{h}_1 \neq \boldsymbol{h}_2) = 0$ and we define the Hilbert space $\mathcal{H}$ with respect to these equivalence classes. The completeness of $\mathcal{H}$ follows from the L_2-completeness theorem (see Loève 1963, p. 161). The space $\mathcal{H}$, equipped with the covariance inner product, thus defines the Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$. Note that for $\boldsymbol{h}_i = (h_{i1}, \ldots, h_{iq})^T$ $(i = 1,2)$, orthogonality ($\boldsymbol{h}_1 \perp \boldsymbol{h}_2$) merely means that $E(\boldsymbol{h}_1^T \boldsymbol{h}_2) = \sum_{j=1}^{q} E(h_{1j} h_{2j}) = 0$. This is different from being uncorrelated, which is stronger and means that

$$E(\boldsymbol{h}_1 \boldsymbol{h}_2^T) = \begin{bmatrix} E(h_{11}h_{21}) & \cdots & E(h_{11}h_{2q}) \\ \vdots & \ddots & \vdots \\ E(h_{1q}h_{21}) & \cdots & E(h_{1q}h_{2q}) \end{bmatrix} = \mathbf{0}^{q \times q}.$$

However, $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ being uncorrelated does imply that $\boldsymbol{h}_1 \perp \boldsymbol{h}_2$.

Definition 2.6 (Closed Linear Subspace)**.** *A space $\mathcal{G} \subset \mathcal{H}$ is a **linear subspace** of $\mathcal{H}$ if $\boldsymbol{g}_i \in \mathcal{G}$ and $\lambda_i \in \mathbb{R}$, $i = 1,2$, implies that $\lambda_1 \boldsymbol{g}_1 + \lambda_2 \boldsymbol{g}_2 \in \mathcal{G}$. A linear subspace is called **closed** if it contains all its limit points.*

An important theorem in the geometry of Hilbert spaces is the Projection Theorem, which is crucial in the development of the semiparametric theory.

Theorem 2.3 (Projection Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{G}$ a closed linear subspace. For any $\boldsymbol{h} \in \mathcal{H}$, there exists a unique $\boldsymbol{g}_0 \in \mathcal{G}$ such that the distance to $\boldsymbol{h}$ is minimized; that is $\|\boldsymbol{h} - \boldsymbol{g}_0\| \leq \|\boldsymbol{h} - \boldsymbol{g}\|$ for all $\boldsymbol{g} \in \mathcal{G}$. Furthermore, $\boldsymbol{h} - \boldsymbol{g}_0$ is orthogonal to $\mathcal{G}$; that is, $\langle \boldsymbol{h} - \boldsymbol{g}_0, \boldsymbol{g} \rangle = 0$ for all $\boldsymbol{g} \in \mathcal{G}$. This is denoted $\boldsymbol{h} - \boldsymbol{g}_0 \perp \mathcal{G}$.*

A proof can be found in Luenberger (1969), Theorem 2, p. 51.

Definition 2.7 (Orthogonal Projection)**.** *The unique element $\boldsymbol{g}_0 \in \mathcal{G}$, constructed in Theorem 2.3, is called the **orthogonal projection** of $\boldsymbol{h}$ onto the space $\mathcal{G}$, and is denoted by $\Pi(\boldsymbol{h}|\mathcal{G})$.*

Note that the Projection Theorem merely guarantees the existence and uniqueness of the orthogonal projection of $\boldsymbol{h} \in \mathcal{H}$ onto the closed linear subspace $\mathcal{G}$, it does not clarify how to construct this projection $\Pi(\boldsymbol{h}|\mathcal{G})$. Fortunately, in some interesting cases, it is not difficult to find an explicit expression for this projection, as the following example will indicate.

We will often be interested in the orthogonal projection of a mean-zero q-dimensional measurable random function $\boldsymbol{h} \in \mathcal{H}$ onto the linear subspace $\mathcal{G}$ spanned by a mean-zero r-dimensional measurable random function (with finite second moments) $\boldsymbol{g}$; that is

$$\mathcal{G} = \left\{ \boldsymbol{B}^{q \times r} \boldsymbol{g} \middle| \boldsymbol{B}^{q \times r} \in \mathbb{R}^{q \times r} \right\} \subset \mathcal{H},$$

with $\mathbb{R}^{q \times r}$ the set of all real matrices with q rows and r columns, for positive integers q and r. In this case, the orthogonal projection of $\boldsymbol{h}$ onto $\mathcal{G}$ is given by the following proposition.

Proposition 2.1. *The orthogonal projection of any q-dimensional random function $\boldsymbol{h} \in \mathcal{H}$ onto the linear subspace $\mathcal{G}$ spanned by the r-dimensional random function $\boldsymbol{g}$ is given by*

$$\Pi(\boldsymbol{h}|\mathcal{G}) = E(\boldsymbol{h}\boldsymbol{g}^T)\{E(\boldsymbol{g}\boldsymbol{g}^T)\}^{-1}\boldsymbol{g}, \tag{2.6}$$

assuming $\boldsymbol{g}$ is as such that $E(\boldsymbol{g}\boldsymbol{g}^T)$ is positive definite.

We end this section with some final useful concepts regarding the geometry of Hilbert spaces.

Definition 2.8 (Direct Sum)**.** *We say that $\mathcal{G}_1 \oplus \mathcal{G}_2$ is a **direct sum** of two linear subspaces $\mathcal{G}_i \subset \mathcal{H}$, $i = 1,2$, if $\mathcal{G}_1 \oplus \mathcal{G}_2$ is a linear subspace of the Hilbert space $\mathcal{H}$ and if every element $\boldsymbol{h} \in \mathcal{G}_1 \oplus \mathcal{G}_2$ has a unique representation of the form $\boldsymbol{h} = \boldsymbol{g}_1 + \boldsymbol{g}_2$, where $\boldsymbol{g}_i \in \mathcal{G}_i$, $i = 1,2$.*

Definition 2.9 (Orthogonal Complement)**.** *Let $\mathcal{G}$ be a linear subspace of $\mathcal{H}$. The **orthogonal complement** of $\mathcal{G}$, denoted by $\mathcal{G}^\perp$, is the linear subspace defined as*

$$\mathcal{G}^\perp = \{\boldsymbol{h} \in \mathcal{H} | \langle \boldsymbol{h}, \boldsymbol{g} \rangle = 0 \text{ for all } \boldsymbol{g} \in \mathcal{G}\}.$$

This allows for the decomposition $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^\perp$. For any $\boldsymbol{h} \in \mathcal{H}$ and a linear subspace $\mathcal{G}$, we have that $\boldsymbol{h} = \Pi(\boldsymbol{h}|\mathcal{G}) + \Pi(\boldsymbol{h}|\mathcal{G}^\perp)$, where $\Pi(\boldsymbol{h}|\mathcal{G})$ is the orthogonal projection of $\boldsymbol{h}$ onto the space $\mathcal{G}$ and $\Pi(\boldsymbol{h}|\mathcal{G}^\perp) = \boldsymbol{h} - \Pi(\boldsymbol{h}|\mathcal{G})$ is the **residual** of $\boldsymbol{h}$ after projecting it onto $\mathcal{G}$.

Definition 2.10 (Linear Variety)**.** *A **linear variety** is the translation of a linear subspace away from the origin; that is, a linear variety $\mathcal{V}$ can be written as $\mathcal{V} = \boldsymbol{h}_0 + \mathcal{G}$, for a linear subspace $\mathcal{G}$ and $\boldsymbol{h}_0 \in \mathcal{H} \setminus \mathcal{G}$ with $\|\boldsymbol{h}_0\| \neq 0$.*

2.3 Geometry of Influence Functions of Parametric Models

To introduce the geometry of influence functions, we start by studying finite-dimensional parametric models. In Section 2.4, these results will be generalized to infinite-dimensional semiparametric models. The content of this section is primarily based on Tsiatis (2006, chap. 3).

Reconsider the i.i.d. random sample $\mathcal{O} = \{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$ for which we believe the true density function $f_{\boldsymbol{O},0}(\boldsymbol{o})$ of a single $\boldsymbol{O}$ belongs to the parametric model $\mathcal{M}_\text{P} = \{f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\theta}) \in \mathcal{M}_\text{NP} | \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$, with $\boldsymbol{\theta}_0$ is such that $f_{\boldsymbol{O},0}(\boldsymbol{o}) = f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\theta}_0)$. Suppose interest lies in the estimation of the q-dimensional parameter $\boldsymbol{\beta}$. In many

cases, the model allows for a parametrization $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$ with an r-dimensional nuisance parameter $\boldsymbol{\eta}$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\eta}_0^T)^T$. This will be the main focus of this section. However, one must keep in mind that some problems lend themselves more naturally to define the parameter of interest as $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta})$, a smooth q-dimensional function of the model parameters $\boldsymbol{\theta}$, with $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\boldsymbol{\theta}_0)$. We will briefly discuss this near the end of this section. This representation is often useful in infinite-dimensional semiparametric models, which we will demonstrate in Section 2.4. An important example will be studied in detail in Section 2.5, namely, the estimation of a mean outcome in the presence of incomplete data, explainable by measured auxiliary covariates.

2.3.1 Regular and asymptotically linear estimators

Let $\hat{\boldsymbol{\beta}}_n$ denote an estimator of $\boldsymbol{\beta}$, which can be seen as a q-dimensional measurable random function of the observed data $\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n$. Below, we define the class of estimators that will be of main interest.

Definition 2.11 (Asymptotically Linear Estimator, ALE)**.** *An **asymptotically linear estimator (ALE)** $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ is an estimator for which there exists a q-dimensional measurable random function $\boldsymbol{\phi}(\boldsymbol{O})$, such that*

(1) $E\{\boldsymbol{\phi}(\boldsymbol{O})\} = \mathbf{0}$,

(2) $E\{\boldsymbol{\phi}(\boldsymbol{O})\boldsymbol{\phi}^T(\boldsymbol{O})\}$ *is non-singular and* $E\{\boldsymbol{\phi}^T(\boldsymbol{O})\boldsymbol{\phi}(\boldsymbol{O})\} < \infty$,

(3) and the estimator allows for the expansion

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\phi}(\boldsymbol{O}_i) + o_p(1), \tag{2.7}$$

with $o_p(1)$ a term that converges in probability to zero as n goes to infinity.

Note that $\boldsymbol{\phi}(\boldsymbol{O})$ is defined with respect to the true (unknown) distribution function $f_{\boldsymbol{O},0}(\boldsymbol{o})$.

Definition 2.12 (Influence Function)**.** *The function $\boldsymbol{\phi}(\boldsymbol{O}_i)$ in (2.7) is referred to as the ith **influence function** of the ith observation $\boldsymbol{O}_i$ of the estimator $\hat{\boldsymbol{\beta}}_n$; $\boldsymbol{\phi}(\boldsymbol{O})$ is simply referred to as the influence function of the estimator $\hat{\boldsymbol{\beta}}_n$.*

Identification of the influence function of an ALE $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ is crucial because it is sufficient to study the asymptotic properties of $\hat{\boldsymbol{\beta}}_n$. This follows from the central limit theorem (CLT) and Slutsky's theorem:

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N\left\{\mathbf{0}, E(\boldsymbol{\phi}\boldsymbol{\phi}^T)\right\},$$

with $\xrightarrow{d}$ denoting convergence in distribution. Furthermore, an ALE is asymptotically uniquely identified through its influence function.

Theorem 2.4. *An ALE has a unique influence function in the sense that if $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are two influence functions of an ALE, then $P(\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2) = 1$ and thus, $\boldsymbol{\phi}_1$ equals $\boldsymbol{\phi}_2$ with probability one (w.p. 1).*

To avoid **super-efficient** estimators (that is, estimators which are asymptotically unbiased, but may have asymptotic variance smaller than the Cràmer-Rao lower bound for some parameter values (see Tsiatis (2006), Section 3.1, for an example of a super-efficient estimator), we will impose some additional regularity conditions. We will require an estimator to be **regular**, defined in the following sense.

Definition 2.13 (Regular Estimator)**.** *Consider a local data generating process, where, for each n, the data are distributed according to $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_n^T, \boldsymbol{\eta}_n^T)^T$, where $n^{1/2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)$ converges to a constant $\boldsymbol{\lambda}$ so that $\boldsymbol{\theta}_n$ is close to the fixed parameter $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*T}, \boldsymbol{\eta}^{*T})^T$. That is, $\boldsymbol{O}_{1,n}, \ldots, \boldsymbol{O}_{n,n}$ are i.i.d. according to $f_{\boldsymbol{O}}(\boldsymbol{o}; \boldsymbol{\theta}_n)$. An estimator $\hat{\boldsymbol{\beta}}_n(\boldsymbol{O}_{1,n}, \ldots, \boldsymbol{O}_{n,n})$ is said to be **regular** if for each $\boldsymbol{\theta}^*$, $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$ has a limiting distribution that does not depend on the local data generating process.*

In what follows, we will restrict ourselves to Regular Asymptotically Linear (RAL) estimators. Characterization of all RAL estimators for a parameter of interest $\boldsymbol{\beta}$ can be accomplished using the notion of a score vector.

Definition 2.14 (Score Vector)**.** *For $\boldsymbol{O} \sim f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\theta})$, the score vector for a single observation is defined as*

$$\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O};\boldsymbol{\theta}_0) = \left.\frac{\partial \log f_{\boldsymbol{O}}(\boldsymbol{O};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

When $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$, the score vector can be written as $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O};\boldsymbol{\theta}_0) = \left\{\boldsymbol{S}_{\boldsymbol{\beta}}^T(\boldsymbol{O};\boldsymbol{\theta}_0), \boldsymbol{S}_{\boldsymbol{\eta}}^T(\boldsymbol{O};\boldsymbol{\theta}_0)\right\}^T$, where

$$\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\theta}_0) = \left.\frac{\partial \log f_{\boldsymbol{O}}(\boldsymbol{O};\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \text{ and } \boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{O};\boldsymbol{\theta}_0) = \left.\frac{\partial \log f_{\boldsymbol{O}}(\boldsymbol{O};\boldsymbol{\theta})}{\partial \boldsymbol{\eta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Proposition 2.2 (Mean Zero Property Score Vector)**.** *Under suitable regularity conditions, we have that $E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O};\boldsymbol{\theta}_0)\} = \boldsymbol{0}$, so that the score function is an unbiased estimating function for the parameter $\boldsymbol{\theta}$.*

We are now ready to give the powerful result that allows us to describe the geometry of influence functions for RAL estimators. First, the result is given for the general representation where the q-dimensional parameter of interest, $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta})$, is a smooth function of the p-dimensional $\boldsymbol{\theta}$. Afterwards, the result is given for the case $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$.

Theorem 2.5. *Let the parameter of interest $\boldsymbol{\beta}(\boldsymbol{\theta})$ be a q-dimensional function of the p-dimensional parameter $\boldsymbol{\theta}$ ($q < p$), such that $\Gamma(\boldsymbol{\theta}) = \partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}^T$ exists, has rank q and is continuous in $\boldsymbol{\theta}$ in a neighborhood of the truth $\boldsymbol{\theta}_0$. Let $\hat{\boldsymbol{\beta}}_n$ be an ALE with influence function $\boldsymbol{\phi}(\boldsymbol{O})$ such that $E_{\boldsymbol{\theta}}(\boldsymbol{\phi}^T\boldsymbol{\phi})$ exists and is continuous in a neighborhood of $\boldsymbol{\theta}_0$. Then, $\hat{\boldsymbol{\beta}}_n$ is RAL if and only if*

$$E\{\boldsymbol{\phi}(\boldsymbol{O})\boldsymbol{S}_{\boldsymbol{\theta}}^T(\boldsymbol{O};\boldsymbol{\theta}_0)\} = \Gamma(\boldsymbol{\theta}_0). \tag{2.8}$$

Corollary 2.1. *If $\boldsymbol{\theta}$ can be partitioned as $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$, we additionally obtain (under the conditions of Theorem 2.5)*

(1) $E\{\boldsymbol{\phi}(\boldsymbol{O})\boldsymbol{S}_{\boldsymbol{\beta}}^T(\boldsymbol{O};\boldsymbol{\theta}_0)\} = \boldsymbol{I}$, and

(2) $E\{\boldsymbol{\phi}(\boldsymbol{O})\boldsymbol{S}_{\boldsymbol{\eta}}^T(\boldsymbol{O};\boldsymbol{\theta}_0)\} = \mathbf{0}$.

For an outline of a proof, we refer to Tsiatis (2006, chap. 3). For a more general proof, we refer to Newey (1990, p. 127-128). Theorem 2.5 and Corollary 2.1 characterize the set $\mathcal{V}$ of all RAL estimators: any RAL estimator has influence function $\boldsymbol{\phi}$ satisfying (2.8) and, conversely, any $\boldsymbol{\phi} \in \mathcal{H}$ satisfying (2.8) is the influence function of some RAL estimator. The full power of Theorem 2.5 is thus that it lends itself to a geometric representation of RAL estimators, which allows us to compare the efficiency of different RAL estimators and to identify the most efficient one, as we will describe next.

2.3.2 Geometry of influence functions of parametric models

Recall Definition 2.4 of the space $\mathcal{H}$. Define the following subspace of $\mathcal{H}$ implied by the parametric model $\mathcal{M}_{\mathrm{P}}$ (2.2).

Definition 2.15 (Tangent Space)**.** *The finite-dimensional subspace of* $\mathcal{H}$ *spanned by the score vector* $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O};\boldsymbol{\theta}_0)$*, that is,*

$$\mathcal{T} = \{\boldsymbol{B}^{q\times p}\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O};\boldsymbol{\theta}_0) | \boldsymbol{B}^{q\times p} \in \mathbb{R}^{q\times p}\}, \tag{2.9}$$

is called the ***tangent space****.*

By Proposition 2.2, it follows that $\mathcal{T} \subset \mathcal{H}$. If $\boldsymbol{\theta}$ can be partitioned as $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$, we can decompose the tangent space accordingly.

Definition 2.16 (Nuisance Tangent Space)**.** *The finite-dimensional subspace of* $\mathcal{H}$ *spanned by the nuisance score vector* $\boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{O};\boldsymbol{\theta}_0)$*, that is,*

$$\Lambda = \{\boldsymbol{B}^{q\times r}\boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{O};\boldsymbol{\theta}_0) | \boldsymbol{B}^{q\times r} \in \mathbb{R}^{q\times r}\}, \tag{2.10}$$

is called the ***nuisance tangent space****.*

Definition 2.17 (Tangent Space of the Parameter of Interest)**.** *The finite-dimensional subspace of* $\mathcal{H}$ *spanned by the score vector* $\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\theta}_0)$*, that is,*

$$\mathcal{T}_{\boldsymbol{\beta}} = \{\boldsymbol{B}^{q\times q}\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\theta}_0)|\boldsymbol{B}^{q\times q} \in \mathbb{R}^{q\times q}\}, \tag{2.11}$$

is called the ***tangent space of the parameter of interest****.*

When this partition of $\boldsymbol{\theta}$ is possible, the tangent space can be written as the direct sum $\mathcal{T} = \mathcal{T}_{\boldsymbol{\beta}} \oplus \Lambda$. From Corollary 2.1, condition (2), it follows that $\boldsymbol{\phi} \perp \Lambda$, i.e., $\boldsymbol{\phi} \in \Lambda^{\perp}$, for any influence function $\boldsymbol{\phi}$. The following result identifies the set of all influence functions, which easily follows from Theorem 2.5.

Theorem 2.6 (Linear Variety of Influence Functions)**.** *The set of all influence functions, namely the elements of $\mathcal{H}$ that satisfy condition (2.8) of Theorem 2.5, is the linear variety*

$$\mathcal{V} = \boldsymbol{\phi}^*(\boldsymbol{O}) + \mathcal{T}^{\perp}, \tag{2.12}$$

where $\boldsymbol{\phi}^*(\boldsymbol{O})$ is an arbitrary influence function and $\mathcal{T}^{\perp}$ is the orthogonal complement of the tangent space $\mathcal{T}$.

2.3.3 The efficient influence function

In Section 2.3.2, we identified the class of all RAL estimators for a parametric model $\mathcal{M}_{\mathrm{P}}$ by identifying the linear variety $\mathcal{V}$ of all influence functions $\boldsymbol{\phi}$. The aim is now to find the RAL estimator with *smallest* asymptotic variance. Because the asymptotic distribution of a RAL estimator is fully characterized by its influence function and the asymptotic variance is given by the variance of this influence function, the aim is thus to find the influence function within $\mathcal{V}$ with smallest variance, referred to as the **efficient influence function**. First, we need to define what is meant by *smaller* in multiple dimensions.

Consider two RAL estimators of $\boldsymbol{\beta}$ with influence functions $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, respectively. We say that

$$\mathrm{var}(\boldsymbol{\phi}_1) \leq \mathrm{var}(\boldsymbol{\phi}_2) \text{ if and only if } \mathrm{var}(\boldsymbol{\lambda}^T\boldsymbol{\phi}_1) \leq \mathrm{var}(\boldsymbol{\lambda}^T\boldsymbol{\phi}_2), \text{ for all } \boldsymbol{\lambda} \in \mathbb{R}^q,$$

or thus, $E(\boldsymbol{\phi}_2\boldsymbol{\phi}_2^T) - E(\boldsymbol{\phi}_1\boldsymbol{\phi}_1^T)$ is nonnegative definite.

For $q = 1$, orthogonality in $\mathcal{H}$ ($h_1 \perp h_2$) implies that $\mathrm{var}(h_1 + h_2) = \mathrm{var}(h_1) +$

var(h_2). This is not necessarily true when $q \geq 2$, since then orthogonality of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ does not imply that $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are uncorrelated. However, there is an important special case when this does occur.

Definition 2.18 (q-Replicating Linear Space)**.** *A linear subspace $\mathcal{G} \subset \mathcal{H}$ is a **q-replicating linear space** if $\mathcal{G}$ is of the form $\mathcal{G} = \mathcal{G}^{(1)} \times \ldots \times \mathcal{G}^{(1)} = \{\mathcal{G}^{(1)}\}^q$, where $\mathcal{G}^{(1)}$ denotes a linear subspace of the Hilbert space of one-dimensional mean-zero random functions of $\boldsymbol{O}$ and $\{\mathcal{G}^{(1)}\}^q$ represents the linear subspace in $\mathcal{H}$ consisting of elements $\boldsymbol{h} = (h^{(1)}, \ldots, h^{(q)})^T$, such that $h^{(j)} \in \mathcal{G}^{(1)}$ for all $j = 1, \ldots, q$.*

By construction, the subspaces $\mathcal{T}$, Λ and $\mathcal{T}_{\boldsymbol{\beta}}$ are examples of such q-replicating linear spaces. For such q-replicating linear spaces, we can extend the Pythagorean Theorem.

Theorem 2.7 (Multivariate Pythagorean Theorem)**.** *For $\boldsymbol{g} \in \mathcal{G} \subset \mathcal{H}$ and $\mathcal{G}$ a q-replicating linear space and for $\boldsymbol{h} \in \mathcal{H}$ such that $\boldsymbol{h} \in \mathcal{G}^{\perp}$, we have that $var(\boldsymbol{g} + \boldsymbol{h}) = var(\boldsymbol{g}) + var(\boldsymbol{h})$. This implies that for any $\boldsymbol{h}^* \in \mathcal{H}$, $var(\boldsymbol{h}^*) = var\{\Pi(\boldsymbol{h}^*|\mathcal{G})\} + var\{\Pi(\boldsymbol{h}^*|\mathcal{G}^{\perp})\}$.*

It is now straightforward to obtain the **efficient influence function**, that is, the influence function with smallest variance, denoted by $\boldsymbol{\phi}_{\text{eff}}(\boldsymbol{O})$. Let $\boldsymbol{\phi} \in \mathcal{V}$ be an arbitrary influence function. We can write $\boldsymbol{\phi} = \Pi(\boldsymbol{\phi}|\mathcal{T}) + \Pi(\boldsymbol{\phi}|\mathcal{T}^{\perp})$. Let $\boldsymbol{\phi}_{\text{eff}} = \Pi(\boldsymbol{\phi}|\mathcal{T})$, by construction $\boldsymbol{\phi}_{\text{eff}} \in \mathcal{V}$. Furthermore, any $\boldsymbol{\phi} \in \mathcal{V}$ can be written as $\boldsymbol{\phi} = \boldsymbol{\phi}_{\text{eff}} + \boldsymbol{\ell}$ for $\boldsymbol{\ell} \in \mathcal{T}^{\perp}$. Hence, because $\mathcal{T}$ is a q-replicating linear space, we find var$(\boldsymbol{\phi}) =$ var$(\boldsymbol{\phi}_{\text{eff}}) +$ var$(\boldsymbol{\ell})$ which implies that var$(\boldsymbol{\phi}_{\text{eff}}) \leq$ var$(\boldsymbol{\phi})$. The following theorem gives an explicit expression of the efficient influence in the parametric model $\mathcal{M}_{\text{P}}$.

Theorem 2.8 (Efficient Influence Function)**.** *The efficient influence function is given by*

$$\begin{aligned} \boldsymbol{\phi}_{eff}(\boldsymbol{O}) &= \boldsymbol{\phi}^*(\boldsymbol{O}) - \Pi\{\boldsymbol{\phi}^*(\boldsymbol{O})|\mathcal{T}^{\perp}\} = \Pi\{\boldsymbol{\phi}^*(\boldsymbol{O})|\mathcal{T}\} && (2.13) \\ &= \Gamma(\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0) \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O}; \boldsymbol{\theta}_0), && (2.14) \end{aligned}$$

with $\boldsymbol{\phi}^*(\boldsymbol{O})$ an arbitrary influence function in the linear variety $\mathcal{V}$ and with $I(\boldsymbol{\theta}_0) = E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{O};\boldsymbol{\theta}_0)\boldsymbol{S}_{\boldsymbol{\theta}}^T(\boldsymbol{O};\boldsymbol{\theta}_0)\}$ the Fisher information matrix.

Note that $\boldsymbol{\phi}_{\text{eff}}$ is the unique influence function in $\mathcal{T}$ and that

$$\text{var}(\boldsymbol{\phi}_{\text{eff}}) = \Gamma(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)\Gamma^T(\boldsymbol{\theta}_0),$$

the Cràmer-Rao lower bound.

In the special case where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$, (2.14) can be written in terms of the **efficient score**.

Definition 2.19 (Efficient Score)**.** *The **efficient score** is defined as the residual of the score vector with respect to the parameter of interest after projecting it onto the nuisance tangent space:*

$$\begin{aligned} \boldsymbol{S}_{eff}(\boldsymbol{O};\boldsymbol{\theta}_0) &= \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\theta}_0) - \Pi\{\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\theta}_0)|\Lambda\} \\ &= \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\theta}_0) - E(\boldsymbol{S}_{\boldsymbol{\beta}}\boldsymbol{S}_{\boldsymbol{\eta}}^T)\{E(\boldsymbol{S}_{\boldsymbol{\eta}}\boldsymbol{S}_{\boldsymbol{\eta}}^T)\}^{-1}\boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{O};\boldsymbol{\theta}_0). \end{aligned}$$

Corollary 2.2. *When the parameter $\boldsymbol{\theta}$ can be partitioned as $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$, the efficient influence function can be written as*

$$\boldsymbol{\phi}_{eff}(\boldsymbol{O};\boldsymbol{\theta}_0) = \{E(\boldsymbol{S}_{eff}\boldsymbol{S}_{eff}^T)\}^{-1}\boldsymbol{S}_{eff}(\boldsymbol{O};\boldsymbol{\theta}_0). \tag{2.15}$$

Note that $\text{var}(\boldsymbol{\phi}_{\text{eff}}) = \{E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^T)\}^{-1}$, the parametric efficiency bound.

2.4 Extension to Semiparametric Models

In this section, we will extend the geometry of influence functions from parametric models $\mathcal{M}_{\text{P}}$ (2.2) to semiparametric models $\mathcal{M}_{\text{SP}}$ (2.3). First, we extend the geometry of influence functions to semiparametric models $\mathcal{M}_{\text{SP}}$, for which the infinite-dimensional $\boldsymbol{\theta}$ can be partitioned as $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$. Afterwards, we give this extension for semiparametric models $\mathcal{M}_{\text{SP}}$, where the parameter of interest $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta})$ is a smooth functional of the infinite-dimensional parameter $\boldsymbol{\theta}$. The content of this section is primarily based on Tsiatis (2006, chap. 4).

Again, we consider the i.i.d. random sample $\mathcal{O} = \{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$ for which we now believe the true density function $f_{\boldsymbol{O},0}(\boldsymbol{o})$ of a single $\boldsymbol{O}$ belongs to the semiparametric model

$$\mathcal{M}_{\text{SP}} = \{f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\beta},\boldsymbol{\eta}) \in \mathcal{M}_{\text{NP}} | \boldsymbol{\beta} \in \Theta_{\boldsymbol{\beta}} \subset \mathbb{R}^q, \\ \boldsymbol{\eta} \in \Theta_{\boldsymbol{\eta}}, \Theta_{\boldsymbol{\eta}} \text{ an infinite-dimensional set}\},$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\eta}_0^T)^T$ is such that $f_{\boldsymbol{O},0}(\boldsymbol{o}) = f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)$. The extension of the geometry of influence functions to infinite-dimensional statistical models is based on the notion of parametric submodels. We implicitly assume that these parametric submodels are **regular** in the sense of Definition A.1 of Newey (1990).

Definition 2.20 (Parametric Submodel)**.** *A **parametric submodel** $\mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}}$ is a class of densities $f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\beta},\boldsymbol{\gamma})$ indexed by the finite-dimensional parameter $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ such that (1) $\mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}} \subset \mathcal{M}_{SP}$ and (2) $f_{\boldsymbol{O},0}(\boldsymbol{o}) \in \mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}}$.*

Definition 2.21 (Semiparametric RAL Estimator)**.** *An estimator for $\boldsymbol{\beta}$ is RAL for $\mathcal{M}_{SP}$ if it is RAL for every parametric submodel $\mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}}$.*

This implies that the class of all influence functions of RAL estimators for $\boldsymbol{\beta}$ for the semiparametric model $\mathcal{M}_{\text{SP}}$ belongs to the class of influence functions of RAL estimators for $\boldsymbol{\beta}$ for any parametric submodel $\mathcal{M}^{\text{P}}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. Consequently, the influence function of a semiparametric RAL estimator must be orthogonal to all parametric submodel nuisance tangent spaces and its variance must be greater than or equal to

$$\sup_{\mathcal{M}^{\text{P}}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \subset \mathcal{M}_{\text{SP}}} \left\{ E\left(\boldsymbol{S}^{\text{eff},T}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \boldsymbol{S}^{\text{eff}}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \right) \right\}^{-1}. \tag{2.16}$$

Definition 2.22 (Semiparametric Efficiency Bound)**.** *The supremum (2.16) is defined to be the **semiparametric efficiency bound**.*

Definition 2.23 (Locally Efficient RAL estimator)**.** *A semiparametric RAL estimator with asymptotic variance achieving (2.16) for $f_{\boldsymbol{O},0}(\boldsymbol{o})$ is said to be **locally efficient** at $f_{\boldsymbol{O},0}(\boldsymbol{o})$. This means that efficiency is attained only locally, at the true $f_{\boldsymbol{O},0}(\boldsymbol{o})$.*

When the complexity of the parametric submodels is increased so that these approach the semiparametric model, the nuisance tangent space expands accordingly, approaching the semiparametric nuisance tangent space. This is formalized in the following definition.

Definition 2.24 (Semiparametric Nuisance Tangent Space)**.** *The **(semiparametric) nuisance tangent space** Λ is defined as the mean-square closure of all parametric submodel nuisance tangent spaces $\Lambda_{\boldsymbol{\gamma}} = \{\boldsymbol{B}^{q\times r}\boldsymbol{S}_{\boldsymbol{\gamma}}(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)|\boldsymbol{B}^{q\times r}\in\mathbb{R}^{q\times r})\}$ for parametric submodels $\mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. More specifically, Λ is the space of all functions $\boldsymbol{h}\in\mathcal{H}$ such that there exists a sequence $\{\boldsymbol{B}_j^{q\times r_j}\boldsymbol{S}_{\boldsymbol{\gamma}_j}\}_{j=1}^{\infty}$ for which $\|\boldsymbol{h}-\boldsymbol{B}_j\boldsymbol{S}_{\boldsymbol{\gamma}_j}\|^2\to 0$ as $j\to\infty$, for a sequence of parametric submodels $\mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}_j}$. Thus,*

$$\Lambda = \overline{\bigcup_{\mathcal{M}^P_{\boldsymbol{\beta},\boldsymbol{\gamma}}\subset\mathcal{M}_{SP}} \Lambda_{\boldsymbol{\gamma}}}.$$

Given the nuisance tangent space Λ, which we assume to be linear as is the case in most applications, we can define the semiparametric efficient score vector and the efficient influence function.

Definition 2.25 (Semiparametric Efficient Score)**.** *The **semiparametric efficient score** for $\boldsymbol{\beta}$ is defined as*

$$\boldsymbol{S}_{eff}(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) = \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) - \Pi\{\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)|\Lambda\}.$$

Theorem 2.9 (Semiparametric Efficiency Bound)**.** *The semiparametric efficiency bound (2.16) equals $\{E(\boldsymbol{S}_{eff}\boldsymbol{S}_{eff}^T)\}^{-1}$, the inverse of the variance of the semiparametric efficient score.*

Definition 2.26 (Efficient Influence Function)**.** *The **efficient influence function** is defined as the influence function of a semiparametric RAL estimator, if it exists, that achieves the semiparametric efficiency bound.*

The following theorem characterizes all influence functions and the efficient influence function in semiparametric models.

Theorem 2.10. *Any semiparametric RAL estimator for $\boldsymbol{\beta}$ must have an influence function $\boldsymbol{\phi}(\boldsymbol{O})$ that satisfies*

(1) $E\{\boldsymbol{\phi}(\boldsymbol{O})\boldsymbol{S}_{\boldsymbol{\beta}}^T(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\} = E\{\boldsymbol{\phi}(\boldsymbol{O})\boldsymbol{S}_{eff}^T(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\} = \boldsymbol{I}$ *and*

(2) $\boldsymbol{\phi}(\boldsymbol{O}) \perp \Lambda$.

Furthermore, the efficient influence function is the unique element satisfying (1) and (2) whose variance equals the semiparametric efficiency bound and is equal to

$$\boldsymbol{\phi}_{eff}(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) = \{E(\boldsymbol{S}_{eff}\boldsymbol{S}_{eff}^T)\}^{-1}\boldsymbol{S}_{eff}(\boldsymbol{O};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0). \tag{2.17}$$

The following theorem gives an expression for the efficient influence function for a semiparametric model $\mathcal{M}_{\text{SP}} = \{f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\theta}) \in \mathcal{M}_{\text{NP}} | \boldsymbol{\theta} \in \Theta, \Theta \text{ an infinite-dimensional set}\}$ where the parameter of interest $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta})$ is a smooth functional of the infinite-dimensional parameter $\boldsymbol{\theta}$.

Theorem 2.11. *If a semiparametric RAL estimator for $\boldsymbol{\beta}$ exists, then the influence function of this estimator must belong to the linear variety $\mathcal{V} = \boldsymbol{\phi}(\boldsymbol{O}) + \mathcal{T}^{\perp}$, where $\boldsymbol{\phi}(\boldsymbol{O})$ is the influence function of any semiparametric RAL estimator for $\boldsymbol{\beta}$ and $\mathcal{T}$ is the semiparametric tangent space (defined as the mean-square closure of all parametric submodel tangent spaces). If a RAL estimator exists that achieves the semiparametric efficiency bound, then the influence function of this estimator is the efficient influence function*

$$\boldsymbol{\phi}_{eff}(\boldsymbol{O}) = \boldsymbol{\phi}(\boldsymbol{O}) - \Pi\{\boldsymbol{\phi}(\boldsymbol{O})|\mathcal{T}^{\perp}\} = \Pi\{\boldsymbol{\phi}(\boldsymbol{O})|\mathcal{T}\}. \tag{2.18}$$

We end this section with a general result which will be useful for later derivations. It is based on the factorization theorem of joint density functions. Sup-

pose $\boldsymbol{O}$ is an m-dimensional random vector $\boldsymbol{O} = (O^{(1)}, \ldots, O^{(m)})^T$ and let $\overline{\boldsymbol{O}}^{(j)} = (O^{(1)}, \ldots, O^{(j)})^T$, for $j = 1, \ldots, m$. From the factorization theorem of joint density functions, it follows that

$$f_{\boldsymbol{O}}(\boldsymbol{o}) = f_{O^{(1)}}(o^{(1)}) \prod_{j=2}^{m} f_{O^{(j)}|\overline{\boldsymbol{O}}^{(j-1)}}(o^{(j)}|\overline{\boldsymbol{o}}^{(j-1)}).$$

This factorization of the joint density function of $\boldsymbol{O}$ into conditional density functions implies an orthogonal decomposition of the Hilbert space $\mathcal{H}$.

Theorem 2.12 (Tangent Space of a Nonparametric Model)**.** *The tangent space $\mathcal{T}$ for the nonparametric model $\mathcal{M}_{NP}$ (2.1) equals the entire Hilbert space $\mathcal{H}$. The Hilbert space can be decomposed as $\mathcal{H} = \mathcal{T} = \mathcal{T}_1 \oplus \ldots \oplus \mathcal{T}_m$, where*

$$\begin{aligned} \mathcal{T}_1 &= \left\{ \boldsymbol{\alpha}_1(O^{(1)}) \in \mathcal{H} \middle| E\left\{ \boldsymbol{\alpha}_1(O^{(1)}) \right\} = \boldsymbol{0} \right\}, \\ \mathcal{T}_j &= \left\{ \boldsymbol{\alpha}_j(\overline{\boldsymbol{O}}^{(j)}) \in \mathcal{H} \middle| E\left\{ \boldsymbol{\alpha}_j(\overline{\boldsymbol{O}}^{(j)}) \middle| \overline{\boldsymbol{O}}^{(j-1)} \right\} = \boldsymbol{0} \right\} \end{aligned}$$

for $j = 2, \ldots, m$. The linear space $\mathcal{T}_j$ can be equivalently written as

$$\left\{ \boldsymbol{h}_{*j}(\overline{\boldsymbol{O}}^{(j)}) - E\left\{ \boldsymbol{h}_{*j}(\overline{\boldsymbol{O}}^{(j)}) \middle| \overline{\boldsymbol{O}}^{(j-1)} \right\} \middle| \boldsymbol{h}_{*j}(\overline{\boldsymbol{O}}^{(j)}) \text{ an arbitrary square-integrable function} \right\}.$$

Furthermore, the subspaces $\mathcal{T}_j$, $j = 1, \ldots, m$ are mutually orthogonal spaces; that is, $\mathcal{T}_{j_1} \perp \mathcal{T}_{j_2}$ for any $j_1 \neq j_2$. Finally, any element $\boldsymbol{h}(\boldsymbol{O}) \in \mathcal{H}$ can be decomposed into orthogonal elements $\boldsymbol{h} = \sum_{j=1}^{m} \boldsymbol{h}_j$, where

$$\begin{aligned} \boldsymbol{h}_1(O^{(1)}) &= E\left\{ \boldsymbol{h}(\boldsymbol{O}) | O^{(1)} \right\}, \\ \boldsymbol{h}_j(\overline{\boldsymbol{O}}^{(j)}) &= E\left\{ \boldsymbol{h}(\boldsymbol{O}) | \overline{\boldsymbol{O}}^{(j)} \right\} - E\left\{ \boldsymbol{h}(\boldsymbol{O}) | \overline{\boldsymbol{O}}^{(j-1)} \right\}, \end{aligned}$$

for $j = 2, \ldots, m$, and note that $\boldsymbol{h}_j(\overline{\boldsymbol{O}}^{(j)}) = \Pi\left\{ \boldsymbol{h}(\boldsymbol{O}) | \mathcal{T}_j \right\}$, for $j = 1, \ldots, m$.

We end this section by noting that $\mathcal{T}_j$ is referred to as the tangent space correspond-

ing to the conditional density function $f_{O^{(j)}|\overline{\boldsymbol{O}}^{(j-1)}}(o^{(j)}|\overline{\boldsymbol{o}}^{(j-1)})$. It consists of all functions $\boldsymbol{\alpha}_j(\overline{\boldsymbol{O}}^{(j)}) \in \mathcal{H}$ of the vector $\overline{\boldsymbol{O}}^{(j)}$ that have conditional mean zero given the vector $\overline{\boldsymbol{O}}^{(j-1)}$.

2.5 Application: Estimation of a Population Mean Outcome with Incomplete Data

We end this chapter with an application of the semiparametric theory (presented in Section 2.4) to the problem of estimating a population mean outcome with incomplete data where the missingness can be explained by a set of measured auxiliary variables. This set-up will be of particular interest in the further development of this thesis. The results presented in this section follow from the general theory in Tsiatis (2006), chap. 6–12, and were originally introduced in Robins et al. (1994); Scharfstein et al. (1999a,b).

Consider a study design which intends to collect i.i.d. data $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ on n individuals drawn from some population of interest, with Y_i a one-dimensional outcome of interest and $\boldsymbol{X}_i$ a set of auxiliary covariates for subject i, say p-dimensional. Interest lies in the estimation of the population mean outcome $\mu_0 = E(Y)$. However, estimation of $E(Y)$ is complicated by the fact that Y_i is not available for all individuals. Let R_i denote the missingness indicator, which codes $R_i = 1$ when Y_i is observed, and $R_i = 0$ when Y_i is missing. The observed data can then be described as the random sample $\mathcal{O} = \{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$ with $\boldsymbol{O}_i = (R_i Y_i, R_i, \boldsymbol{X}_i)$. We assume that the covariates $\boldsymbol{X}_i$ contain sufficient information to explain missingness so that the **missing at random (MAR)** assumption, i.e., $Y_i \perp\!\!\!\perp R_i | \boldsymbol{X}_i$ (Rubin 1976), holds.

2.5.1 The model

The MAR assumption implies that the density function function of a single $\boldsymbol{O}$ can be decomposed as

$$f_{\boldsymbol{O}}(\boldsymbol{o}) = f_{RY,R,\boldsymbol{X}}(ry, r, \boldsymbol{x}) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})^r f_{R|\boldsymbol{X}}(r|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{2.19}$$

Since R is binary, the conditional density function of R given $\boldsymbol{X}$ can be written as $f_{R|\boldsymbol{X}}(r|\boldsymbol{x}) = \pi(\boldsymbol{x})^r\{1-\pi(\boldsymbol{x})\}^{1-r}$, with $\pi(\boldsymbol{x})$ the probability of observing the outcome given covariates $\boldsymbol{X} = \boldsymbol{x}$, and is referred to as the **missingness mechanism**. The truth is denoted by $f_{\boldsymbol{O},0}(\boldsymbol{o}) = f_{Y|\boldsymbol{X},0}(y|\boldsymbol{x})^r\pi_0(\boldsymbol{x})^r\{1-\pi_0(\boldsymbol{x})\}^{1-r}f_{\boldsymbol{X},0}(\boldsymbol{x})$. We assume that $\pi_0(\boldsymbol{X}) \geq \delta > 0$ with probability one. This key assumption in the development of the semiparametric theory is called the **positivity assumption** (van der Laan and Rose 2011, chap. 10). The missingness of the outcome can be by design in which case the probability $\pi_0(\boldsymbol{X})$ is known to the investigator. However, when missingness in the outcome occurs by happenstance, the function $\pi_0(\boldsymbol{X})$ is unknown. In the development below, we will assume that the unknown function $\pi_0(\boldsymbol{X})$ is known to be a function of the form $\pi(\boldsymbol{x};\boldsymbol{\psi})$, known as a function of $\boldsymbol{\psi}$, with $\boldsymbol{\psi}$ an unknown finite-dimensional parameter, say s-dimensional. We define $\boldsymbol{\psi}_0$ (the truth) such that $\pi(\boldsymbol{x};\boldsymbol{\psi}_0) = \pi_0(\boldsymbol{x})$. We conclude with the following semiparametric model:

$$\mathcal{M}_{\text{SP}} = \Big\{ f_{\boldsymbol{O}}(\boldsymbol{o};\boldsymbol{\eta},\boldsymbol{\psi}) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x};\eta_Y)^r\pi(\boldsymbol{x};\boldsymbol{\psi})^r\{1-\pi(\boldsymbol{x};\boldsymbol{\psi})\}^{1-r}f_{\boldsymbol{X}}(\boldsymbol{x};\eta_{\boldsymbol{X}}) \Big| \boldsymbol{\psi} \in \Psi \subset \mathbb{R}^q, \boldsymbol{\eta} = (\eta_Y,\eta_{\boldsymbol{X}}) \Big\}, \tag{2.20}$$

with η_Y and $\eta_{\boldsymbol{X}}$ infinite-dimensional nuisance parameters. Specifically, η_Y indexes the set of all conditional density functions of Y given $\boldsymbol{X}$ and $\eta_{\boldsymbol{X}}$ indexes the set of all joint density functions for the covariates $\boldsymbol{X}$. The truth $(\eta_{Y,0},\eta_{\boldsymbol{X},0})$ is defined such that $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x};\eta_{Y,0}) = f_{Y|\boldsymbol{X},0}(y|\boldsymbol{x})$ and $f_{\boldsymbol{X}}(\boldsymbol{x};\eta_{\boldsymbol{X},0}) = f_{\boldsymbol{X},0}(\boldsymbol{x})$. The parameter of interest is the functional

$$\mu_0 = E\{E(Y|\boldsymbol{X})\} = \int \left\{ \int y f_{Y|\boldsymbol{X},0}(y|\boldsymbol{x})dy \right\} f_{\boldsymbol{X},0}(\boldsymbol{x})d\boldsymbol{x}.$$

2.5.2 Semiparametric efficiency theory

The tangent space $\mathcal{T}$

To construct the linear variety of all influence functions for μ_0, we first identify the tangent space $\mathcal{T}$. Recall that this is the mean-square closure off all parametric submodel tangent spaces. The factorization of the observed data density function into variational independent parts implies a decomposition of the tangent space:

$\mathcal{T} = \mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}} \oplus \Lambda_{\boldsymbol{\psi}}$. The space $\mathcal{T}_Y$ corresponds to the tangent space of the factor $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})^r$. Because the conditional density function $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ is left completely unspecified, it follows that

$$\mathcal{T}_Y = \{R\alpha_Y(Y,\boldsymbol{X})|E\{\alpha_Y(Y,\boldsymbol{X})|\boldsymbol{X}\} = 0\}. \tag{2.21}$$

Similarly, because the density function $f_{\boldsymbol{X}}(\boldsymbol{x})$ is also left completely unspecified, we have that

$$\mathcal{T}_{\boldsymbol{X}} = \{\alpha_{\boldsymbol{X}}(\boldsymbol{X})|E\{\alpha_{\boldsymbol{X}}(\boldsymbol{X})\} = 0\}. \tag{2.22}$$

Finally, $\Lambda_{\boldsymbol{\psi}}$ is the tangent space of the parametric model for the missingness mechanism implied by $\pi(\boldsymbol{X};\boldsymbol{\psi})$ and is spanned by the score vector for $\boldsymbol{\psi}$:

$$\Lambda_{\boldsymbol{\psi}} = \left\{\boldsymbol{b}^T \boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\middle|\boldsymbol{b} \in \mathbb{R}^s\right\} \tag{2.23}$$

where the score vector for $\boldsymbol{\psi}$ is given by

$$\begin{aligned}\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0) &= \left.\frac{\partial \log\left[\pi(\boldsymbol{X};\boldsymbol{\psi})^R\{1-\pi(\boldsymbol{X};\boldsymbol{\psi})\}^{1-R}\right]}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \\ &= \frac{R-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}}\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0),\end{aligned}$$

with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0) = \partial\pi(\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$. Note that

$$\Lambda_{\boldsymbol{\psi}} \subset \mathcal{T}_R = \{\alpha_R(R,\boldsymbol{X})|E\{\alpha_R(R,\boldsymbol{X})|\boldsymbol{X}\} = 0\},$$

with $\mathcal{T}_R$ the tangent space of a nonparametric model for the missingness mechanism.

Now take arbitrary functions $R\alpha_Y(Y,\boldsymbol{X}) \in \mathcal{T}_Y$, $\alpha_{\boldsymbol{X}}(\boldsymbol{X}) \in \mathcal{T}_{\boldsymbol{X}}$ and $\alpha_R(Y,\boldsymbol{X}) \in \mathcal{T}_R$. First observe that $E\{\alpha_R(R,\boldsymbol{X})\alpha_{\boldsymbol{X}}(\boldsymbol{X})\} = E[E\{\alpha_R(R,\boldsymbol{X})|\boldsymbol{X}\}\alpha_{\boldsymbol{X}}(\boldsymbol{X})] = 0$. Next, observe that $E\{\alpha_R(R,\boldsymbol{X})R\alpha_Y(Y,\boldsymbol{X})\} = E[\alpha_R(R,\boldsymbol{X})RE\{\alpha_Y(Y,\boldsymbol{X})|R,\boldsymbol{X}\}]$, which, by the MAR assumption, equals $E[\alpha_R(R,\boldsymbol{X})RE\{\alpha_Y(Y,\boldsymbol{X})|\boldsymbol{X}\}] = 0$. Finally, we have that $E\{\alpha_{\boldsymbol{X}}(\boldsymbol{X})R\alpha_Y(Y,\boldsymbol{X})\} = E[\alpha_{\boldsymbol{X}}(\boldsymbol{X})RE\{\alpha_Y(Y,\boldsymbol{X})|R,\boldsymbol{X}\}]$, which equals, also by the MAR assumption, $E[\alpha_{\boldsymbol{X}}(\boldsymbol{X})RE\{\alpha_Y(Y,\boldsymbol{X})|\boldsymbol{X}\}] = 0$. From this, it follows that $\mathcal{T}_Y$, $\mathcal{T}_{\boldsymbol{X}}$ and $\Lambda_{\boldsymbol{\psi}}$ are mutually orthogonal spaces.

We conclude with the following result.

Proposition 2.3 (Tangent Space $\mathcal{T}$)**.** *The tangent space of the semiparametric model $\mathcal{M}_{SP}$ (2.20) is given by*

$$\begin{aligned}\mathcal{T} &= \mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}} \oplus \Lambda_{\boldsymbol{\psi}} \\ &= \{R\alpha_Y(Y,\boldsymbol{X}) + \alpha_{\boldsymbol{X}}(\boldsymbol{X}) + \boldsymbol{b}^T \boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)| \\ &\qquad R\alpha_Y(Y,\boldsymbol{X}) \in \mathcal{T}_Y, \alpha_{\boldsymbol{X}}(\boldsymbol{X}) \in \mathcal{T}_{\boldsymbol{X}}, \boldsymbol{b} \in \mathbb{R}^s\},\end{aligned} \tag{2.24}$$

with $\Lambda_{\boldsymbol{\psi}} \perp \mathcal{T}_{\boldsymbol{X}}$, $\Lambda_{\boldsymbol{\psi}} \perp \mathcal{T}_Y$, and $\mathcal{T}_{\boldsymbol{X}} \perp \mathcal{T}_Y$.

The orthogonal complement $\mathcal{T}^\perp$

Since the entire Hilbert space $\mathcal{H}$ corresponding to the observed data vector $\boldsymbol{O} = (RY, R, \boldsymbol{X})$ can be decomposed as $\mathcal{H} = \mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}} \oplus \mathcal{T}_R$ with $\mathcal{T}_Y$, $\mathcal{T}_{\boldsymbol{X}}$ and $\mathcal{T}_R$ mutually orthogonal, it follows that $(\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}})^\perp = \mathcal{T}_R$. Next, because $\mathcal{T} = (\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}}) \oplus \Lambda_{\boldsymbol{\psi}}$, $\Lambda_{\boldsymbol{\psi}} \subset \mathcal{T}_R = (\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}})^\perp$, it follows that $\mathcal{T}^\perp = \mathcal{T}_R \cap \Lambda_{\boldsymbol{\psi}}^\perp$ and thus $\mathcal{T}^\perp = \Pi(\mathcal{T}_R | \Lambda_{\boldsymbol{\psi}}^\perp) = \mathcal{T}_R - \Pi(\mathcal{T}_R | \Lambda_{\boldsymbol{\psi}})$.

A more explicit description of the orthogonal complement of the tangent space can be obtained as follows. Any function in $\mathcal{T}_R$ can be written as $\alpha_R(R,\boldsymbol{X}) = R\alpha(1,\boldsymbol{X}) + (1-R)\alpha(0,\boldsymbol{X})$ and because $E\{\alpha_R(R,\boldsymbol{X})\} = 0$, we find that $\alpha(1,\boldsymbol{X}) = -[\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}/\pi(\boldsymbol{X};\boldsymbol{\psi}_0)]\boldsymbol{\alpha}(0,\boldsymbol{X})$. Consequently, the space $\mathcal{T}_R$ can be described as the space of all functions of the form $\{1 - R/\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}\tilde{\alpha}(\boldsymbol{X})$ with $\tilde{\alpha}(\boldsymbol{X})$ an arbitrary square-integrable function. Finally, from Proposition 2.1, it follows that

$$\begin{aligned}&\Pi\left[\left\{1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X})\middle|\Lambda_{\boldsymbol{\psi}}\right] \\ &\qquad = E\left[\left\{1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X})\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right] \\ &\qquad\qquad \times [E\{\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\}]^{-1}\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0) \\ &\qquad = -E\left[\tilde{\alpha}(\boldsymbol{X})\frac{\pi_{\boldsymbol{\psi}}^T(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right]\left(E\left[\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\pi_{\boldsymbol{\psi}}^T(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}}\right]\right)^{-1} \\ &\qquad\qquad \times \frac{R - \pi(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}}\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0).\end{aligned}$$

We conclude with the following result.

Proposition 2.4 (Orthogonal Complement $\mathcal{T}^{\perp}$)**.** *The orthogonal complement of the tangent space of the semiparametric model $\mathcal{M}_{SP}$ (2.20) is given by $\mathcal{T}^{\perp} = \Pi(\mathcal{T}_R|\Lambda_{\boldsymbol{\psi}}^{\perp}) = \mathcal{T}_R - \Pi(\mathcal{T}_R|\Lambda_{\boldsymbol{\psi}})$, which equals*

$$\left\{ \left\{ 1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} \right\} \tilde{\alpha}(\boldsymbol{X}) - \Pi\left[\left\{ 1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} \right\} \tilde{\alpha}(\boldsymbol{X}) \middle| \Lambda_{\boldsymbol{\psi}} \right] \middle| \; \tilde{\alpha}(\boldsymbol{X}) \text{ an arbitrary square-integrable function} \right\}$$

with the projection given by the expression above.

We end the discussion of the tangent space $\mathcal{T}$ and its orthogonal complement $\mathcal{T}^{\perp}$ by noting that the space $\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}}$ is the tangent space of the semiparametric model with known missingness mechanism $\pi(\boldsymbol{X};\boldsymbol{\psi}_0)$, e.g., when missingness is by design. The corresponding orthogonal complement is given by $(\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}})^{\perp} = \mathcal{T}_R$, which yields a simplification of the space of influence functions. Later on, we will see that the efficient influence function is the same whether or not the missingness mechanism is known.

The space of influence functions $\phi_{\text{IPTW}}(\boldsymbol{O}) + \mathcal{T}^{\perp}$

To find the space of all influence functions, we need to identify the influence function of an arbitrary root-n consistent RAL estimator of μ_0. For this purpose, consider the standard IPTW (inverse probability of treatment weighted) estimator (Horvitz and Thompson 1952) $\hat{\mu}_{n,\text{IPTW}} = n^{-1}\sum_{i=1}^{n} R_i Y_i / \pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)$, with $\hat{\boldsymbol{\psi}}_n$ the MLE of $\boldsymbol{\psi}_0$. From a standard Taylor expansion, it follows that

$$\begin{aligned} & n^{1/2}(\hat{\mu}_{n,\text{IPTW}} - \mu_0) \\ & \quad = n^{-1/2}\sum_{i=1}^{n} \left\{ \frac{R_i Y_i}{\pi(\boldsymbol{X}_i;\boldsymbol{\psi}_0)} - \mu_0 \right\} \\ & \qquad + \left\{ -n^{-1}\sum_{i=1}^{n} \frac{R_i Y_i}{\pi^2(\boldsymbol{X}_i;\tilde{\boldsymbol{\psi}}_n)} \pi_{\boldsymbol{\psi}}^{T}(\boldsymbol{X}_i;\tilde{\boldsymbol{\psi}}_n) \right\} n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0), \end{aligned}$$

where $\tilde{\boldsymbol{\psi}}_n$ is an intermediate value on the line segment connecting $\hat{\boldsymbol{\psi}}_n$ and $\boldsymbol{\psi}_0$. Because $\hat{\boldsymbol{\psi}}_n$ is the MLE of $\boldsymbol{\psi}_0$, $\hat{\boldsymbol{\psi}}_n$ is a root-n consistent RAL estimator of $\boldsymbol{\psi}_0$ and has as influence function the efficient influence function under the parametric model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for the missingness mechanism. That is,

$$\begin{aligned} &n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}[E\{\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\}]^{-1}\boldsymbol{S}_{\boldsymbol{\psi}}(R_i,\boldsymbol{X}_i;\boldsymbol{\psi}_0) + o_p(1). \end{aligned}$$

Furthermore, because $\hat{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}_0$, we also have that $\tilde{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}_0$ and under suitable regularity conditions (Robins et al. 1994, app. B), it follows from the uniform weak law of large numbers that

$$\begin{aligned} -n^{-1}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(\boldsymbol{X}_i;\tilde{\boldsymbol{\psi}}_n)}\pi_{\boldsymbol{\psi}}^T(\boldsymbol{X}_i;\tilde{\boldsymbol{\psi}}_n) &\xrightarrow{p} -E\left\{RY\frac{\pi_{\boldsymbol{\psi}}^T(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\} \\ &= -E\left[\left\{\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0\right\}\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right]. \end{aligned}$$

It follows that

$$\begin{aligned} &n^{1/2}(\hat{\mu}_{n,\mathrm{IPTW}} - \mu_0) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}\phi_{\mathrm{IPTW}}(\boldsymbol{O}_i) + o_p(1) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}\left[\left\{\frac{R_iY_i}{\pi(\boldsymbol{X}_i;\boldsymbol{\psi}_0)} - \mu_0\right\} - \Pi\left\{\frac{R_iY_i}{\pi(\boldsymbol{X}_i;\boldsymbol{\psi}_0)} - \mu_0\middle|\Lambda_{\boldsymbol{\psi}}\right\}\right] + o_p(1). \end{aligned}$$

We conclude with the following result.

Proposition 2.5 (Space of Influence Functions)**.** *The space of all influence functions for μ_0 under model $\mathcal{M}_{SP}$ (2.20) is given by*

$$\mathcal{V} = \left\{\phi_{\tilde{\alpha}}(\boldsymbol{O}) = \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} + \left\{1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X}) - \mu_0 \right. \tag{2.25}$$

$$-\Pi\left[\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}+\left\{1-\frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X})-\mu_0\middle|\Lambda_{\boldsymbol{\psi}}\right]\Bigg| \quad (2.26)$$

$$\tilde{\alpha}(\boldsymbol{X}) \text{ an arbitrary square-integrable function}\Bigg\}.$$

Construction of semiparametric RAL estimators

Now we have derived the space of all influence functions, it is straightforward to construct semiparametric RAL estimators of μ_0. This is accomplished by using the first part (2.25) of the influence function (2.25-2.26) as an estimating function for μ_0. Consider the estimator

$$\hat{\mu}_n(\tilde{\alpha}) = n^{-1}\sum_{i=1}^{n}\left[\frac{R_iY_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)}+\left\{1-\frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)}\right\}\tilde{\alpha}(\boldsymbol{X}_i)\right], \quad (2.27)$$

with $\hat{\boldsymbol{\psi}}_n$ the MLE of $\boldsymbol{\psi}_0$. From a Taylor expansion similar to that of $\hat{\mu}_{n,\text{IPTW}}$, we find that the influence function of the estimator $\hat{\mu}_n(\tilde{\alpha})$ is indeed given by $\phi_{\tilde{\alpha}}(\boldsymbol{O})$ (2.25-2.26). It follows that $\hat{\mu}_n(\tilde{\alpha})$ is asymptotically normal with asymptotic variance equal to the variance of the influence function $\phi_{\tilde{\alpha}}(\boldsymbol{O})$:

$$\begin{aligned}\text{var}\{\phi_{\tilde{\alpha}}(\boldsymbol{O})\} =& \text{var}\left[\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}+\left\{1-\frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X})-\mu_0\right]\\ &-\text{var}\left(\Pi\left[\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}+\left\{1-\frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X})-\mu_0\middle|\Lambda_{\boldsymbol{\psi}}\right]\right),\end{aligned}$$

which follows from the Pythagorean theorem. Note that when the missingness mechanism would be known to us, that is, $\boldsymbol{\psi}_0$ is known and used instead of $\hat{\boldsymbol{\psi}}_n$ in the construction of $\hat{\mu}_n(\tilde{\alpha})$, the influence function of the estimator would be reduced to (2.25). In this case, the asymptotic variance of this estimator based on the true value $\boldsymbol{\psi}_0$ would equal

$$\text{var}\left[\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}+\left\{1-\frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\}\tilde{\alpha}(\boldsymbol{X})-\mu_0\right],$$

which is larger than or equal to $\text{var}\{\phi_{\tilde{\alpha}}(\boldsymbol{O})\}$. This leads to the (perhaps unintuitive) fact that, even if we know the missingness probabilities (e.g., by design), efficiency could be gained by estimating $\boldsymbol{\psi}$ via MLE in a model that contains the truth rather than using $\boldsymbol{\psi}_0$ itself (Pierce 1982; Rotnitzky et al. 2010).

The efficient influence function $\phi_{\text{eff}}(\boldsymbol{O})$

Proposition 2.5 delineates the space of all influence functions $\mathcal{V}$ for μ_0 under model $\mathcal{M}_{\text{SP}}$ (2.20). Furthermore, for any square-integrable function $\tilde{\alpha}(\boldsymbol{X})$, we know that the RAL estimator $\hat{\mu}_n(\tilde{\alpha})$ has influence function given by $\phi_{\tilde{\alpha}}(\boldsymbol{O})$. It now remains to identify the optimal function $\tilde{\alpha}_{\text{eff}}(\boldsymbol{X})$ such that $\hat{\mu}_n(\tilde{\alpha}_{\text{eff}})$ has the smallest asymptotic variance among the class of all semiparametric RAL estimators under model $\mathcal{M}_{\text{SP}}$ (2.20) and thus the semiparametric RAL estimator that has influence function the efficient influence function $\phi_{\text{eff}}(\boldsymbol{O}) = \phi_{\tilde{\alpha}_{\text{eff}}}(\boldsymbol{O})$.

From Theorem 2.11, it follows that the efficient influence function is given by (2.18), the projection of any influence function, e.g., ϕ_{IPTW}, onto the model tangent space $\mathcal{T}$. We thus have that $\phi_{\text{eff}}(\boldsymbol{O}) = \Pi\{\phi_{\text{IPTW}}(\boldsymbol{O})|(\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}}) \oplus \Lambda_{\boldsymbol{\psi}}\}$. Because $\Lambda_{\boldsymbol{\psi}} \perp \mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}}$, the efficient influence function equals $\Pi\{\phi_{\text{IPTW}}(\boldsymbol{O})|\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}}\} + \Pi\{\phi_{\text{IPTW}}(\boldsymbol{O})|\Lambda_{\boldsymbol{\psi}}\}$. Recall that

$$\phi_{\text{IPTW}}(\boldsymbol{O}) = \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 - \Pi\left\{ \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| \Lambda_{\boldsymbol{\psi}} \right\},$$

and thus $\phi_{\text{IPTW}}(\boldsymbol{O}) = \Pi\{RY/\pi(\boldsymbol{X};\boldsymbol{\psi}_0) - \mu_0|\Lambda_{\boldsymbol{\psi}}^{\perp}\}$. Consequently, the projection $\Pi\{\phi_{\text{IPTW}}(\boldsymbol{O})|\Lambda_{\boldsymbol{\psi}}\} = 0$. Next, because $(\mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}})^{\perp} = \mathcal{T}_R$, we see that $\phi_{\text{eff}}(\boldsymbol{O}) = \phi_{\text{IPTW}}(\boldsymbol{O}) - \Pi\{\phi_{\text{IPTW}}(\boldsymbol{O})|\mathcal{T}_R\}$ and thus

$$\begin{aligned}
\phi_{\text{eff}}(\boldsymbol{O}) =& \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 - \Pi\left\{ \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| \mathcal{T}_R \right\} \\
& - \Pi\left\{ \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| \Lambda_{\boldsymbol{\psi}} \right\} + \Pi\left[\Pi\left\{ \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| \Lambda_{\boldsymbol{\psi}} \right\} \middle| \mathcal{T}_R \right] \\
=& \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 - \Pi\left\{ \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| \mathcal{T}_R \right\}
\end{aligned}$$

because $\Lambda_{\boldsymbol{\psi}} \subset \mathcal{T}_R$. Finally, from Theorem 2.12, it follows that the projection

$\Pi\{RY/\pi(\boldsymbol{X};\boldsymbol{\psi}_0) - \mu_0|\mathcal{T}_R\}$ equals

$$E\left\{\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| R,\boldsymbol{X}\right\} - E\left\{\frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \mu_0 \middle| \boldsymbol{X}\right\}$$
$$= -\left\{1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\} E(Y|\boldsymbol{X}).$$

We may conclude with the following result.

Proposition 2.6 (Efficient Influence Function)**.** *The choice* $\tilde{\alpha}_{eff}(\boldsymbol{X}) = E(Y|\boldsymbol{X})$ *delivers the efficient influence function for* μ_0 *under model* $\mathcal{M}_{SP}$ *(2.20):*

$$\phi_{eff}(\boldsymbol{O}) = \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} + \left\{1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right\} E(Y|\boldsymbol{X}) - \mu_0. \tag{2.28}$$

Note that because the projection of $\phi_{\text{eff}}(\boldsymbol{O})$ onto $\Lambda_{\boldsymbol{\psi}}$ is zero, estimating the missingness probabilities does not lead to an efficiency gain as compared to using the true missingness probabilities. Furthermore, this also implies that if $E(Y|\boldsymbol{X})$ would be known to us, no correction for the estimation of the missingness probabilities would be necessary in the calculation of the asymptotic variance of the efficient RAL estimator $\hat{\mu}_n\{E(Y|\boldsymbol{X})\}$ (see later).

2.5.3 Estimation and inference

Construction of a locally efficient estimator

The aim here is to construct a semiparametric RAL estimator whose influence function is given by the efficient influence function $\phi_{\text{eff}}(\boldsymbol{O})$. The estimator (2.27) with the choice $\tilde{\alpha}_{\text{eff}}(\boldsymbol{X}) = E(Y|\boldsymbol{X})$ has influence function $\phi_{\text{eff}}(\boldsymbol{O})$. However, it is infeasible to calculate because the conditional mean outcome $E(Y|\boldsymbol{X})$ is unknown to us. We will therefore need to posit a working model for the conditional mean outcome, say $m(\boldsymbol{X};\boldsymbol{\xi})$, for some finite-dimensional parameter $\boldsymbol{\xi}$. If this working model is correctly specified, we define $\boldsymbol{\xi}_0$ to be such that $m(\boldsymbol{X};\boldsymbol{\xi}_0) = E(Y|\boldsymbol{X})$.

This leads the following algorithm to estimate μ_0 under the semiparametric model $\mathcal{M}_{\text{SP}}$ (2.20):

1. Posit a model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for the missingness mechanism $\pi_0(\boldsymbol{X})$ which we

assume to be correctly specified ($\boldsymbol{\psi}_0$ satisfying $\pi(\boldsymbol{X};\boldsymbol{\psi}_0) = \pi_0(\boldsymbol{X})$ is well-defined). A popular choice would be to use the logistic regression model

$$\pi(\boldsymbol{X};\boldsymbol{\psi}) = \text{expit}\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\} = \frac{\exp\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\}}{1+\exp\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\}}, \tag{2.29}$$

$\boldsymbol{l}(\boldsymbol{X}) = (1, X_1, \ldots, X_p)^T$ (so $s = p+1$). Given the parametric model $\pi(\boldsymbol{X};\boldsymbol{\psi})$, obtain the MLE $\hat{\boldsymbol{\psi}}_n$, thus $\hat{\boldsymbol{\psi}}_n$ solves the score equation $\sum_{i=1}^n \boldsymbol{S}_{\boldsymbol{\psi}}(R_i, \boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n) = \boldsymbol{0}$. For the logistic regression model, $\hat{\boldsymbol{\psi}}_n$ solves

$$\sum_{i=1}^n \left[R_i - \text{expit}\{\hat{\boldsymbol{\psi}}_n^T \boldsymbol{l}(\boldsymbol{X})\}\right] \boldsymbol{l}(\boldsymbol{X}) = \boldsymbol{0}. \tag{2.30}$$

2. Posit a model $m(\boldsymbol{X};\boldsymbol{\xi})$ for the conditional mean outcome $E(Y|\boldsymbol{X})$, which we do not necessarily assume to be correctly specified. A popular choice would be the linear regression model

$$m(\boldsymbol{X};\boldsymbol{\xi}) = \boldsymbol{\xi}^T \boldsymbol{k}(\boldsymbol{X}), \tag{2.31}$$

$\boldsymbol{k}(\boldsymbol{X}) = (1, X_1, \ldots, X_p)$. Given the parametric model $m(\boldsymbol{X};\boldsymbol{\xi})$, obtain an estimator $\hat{\boldsymbol{\xi}}_n$, e.g., the least squares estimator based on the complete cases. That is, $\hat{\boldsymbol{\xi}}_n$ solves

$$\sum_{i=1}^n R_i \left\{Y_i - \hat{\boldsymbol{\xi}}_n^T \boldsymbol{k}(\boldsymbol{X}_i)\right\} \boldsymbol{k}(\boldsymbol{X}_i) = \boldsymbol{0}. \tag{2.32}$$

This is equivalent to the MLE of $\boldsymbol{\xi}$ under the normal-error linear regression model. We will therefore often refer to the least squares estimator as the MLE.

3. The estimator for μ_0 is then obtained as

$$\begin{aligned}&\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) \\ &\quad = n^{-1} \sum_{i=1}^n \left[\frac{R_i Y_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)} + \left\{1 - \frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)}\right\} m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n)\right].\end{aligned} \tag{2.33}$$

Under suitable regularity conditions, the estimator $\hat{\boldsymbol{\xi}}_n$ will converge in prob-

ability to some constant $\boldsymbol{\xi}^*$; i.e., $\hat{\boldsymbol{\xi}}_n \xrightarrow{p} \boldsymbol{\xi}^*$ and $\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^* = O_p(n^{-1/2})$ (that is, $n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*)$ is bounded in probability). Consequently, the function $m(\boldsymbol{X};\hat{\boldsymbol{\xi}}_n)$ will converge in probability to the function $m(\boldsymbol{X};\boldsymbol{\xi}^*)$. When the working model $m(\boldsymbol{X};\boldsymbol{\xi})$ is correctly specified, $\boldsymbol{\xi}^*$ is the value satisfying $E(Y|\boldsymbol{X}) = m(\boldsymbol{X};\boldsymbol{\xi}^*)$ and thus, in this case, $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0$.

Choosing $\tilde{\boldsymbol{\alpha}}(\boldsymbol{X})$ to be $m(\boldsymbol{X};\boldsymbol{\xi}^*)$ and $\hat{\boldsymbol{\psi}}_n$ the MLE of $\boldsymbol{\psi}_0$, we already know that the estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \boldsymbol{\xi}^*) = \hat{\mu}_n\{m(\boldsymbol{X};\boldsymbol{\xi}^*)\}$ has influence function $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})$. We now have that

$$\begin{aligned}
&n^{1/2}\left\{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \mu_0\right\} \\
&\quad = n^{1/2}\left\{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \boldsymbol{\xi}^*)\right\} + n^{1/2}\left\{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \boldsymbol{\xi}^*) - \mu_0\right\} \\
&\quad = n^{1/2}\left\{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \boldsymbol{\xi}^*)\right\} + n^{-1/2}\sum_{i=1}^{n} \phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O}_i) + o_p(1).
\end{aligned}$$

Next, consider the Taylor expansion

$$\begin{aligned}
&\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \boldsymbol{\xi}^*) \\
&\quad = \left[n^{-1}\sum_{i=1}^{n}\left\{1 - \frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)}\right\}\left\{m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n) - m(\boldsymbol{X}_i;\boldsymbol{\xi}^*)\right\}\right] \\
&\quad = \left[n^{-1}\sum_{i=1}^{n}\left\{1 - \frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)}\right\}\frac{\partial m(\boldsymbol{X}_i;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T}\bigg|_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}_n}\right](\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*),
\end{aligned}$$

where $\tilde{\boldsymbol{\xi}}_n$ is an intermediate value on the line segment connecting $\hat{\boldsymbol{\xi}}_n$ and $\boldsymbol{\xi}^*$. Since $\hat{\boldsymbol{\xi}}_n \xrightarrow{p} \boldsymbol{\xi}^*$, we also have that $\tilde{\boldsymbol{\xi}}_n \xrightarrow{p} \boldsymbol{\xi}^*$ and additionally, because also $\hat{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}^*$, it follows from the uniform WLLN (see Newey and McFadden (1994), Lemma 4.3) that under suitable regularity conditions (Robins et al. 1994, app. B)

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\left\{1 - \frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)}\right\}\frac{\partial m(\boldsymbol{X}_i;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T}\bigg|_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}_n} \\
&\qquad \xrightarrow{p} E\left[\left\{1 - \frac{R}{\pi_0(\boldsymbol{X})}\right\}\frac{\partial m(\boldsymbol{X};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}\right].
\end{aligned}$$

Because $\partial m(\boldsymbol{X};\boldsymbol{\xi})/\partial \boldsymbol{\xi}^T|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}$ is a function of $\boldsymbol{X}$ only, this expectation equals

$$E\left[E\left\{1-\frac{R}{\pi_0(\boldsymbol{X})}\middle|\boldsymbol{X}\right\}\frac{\partial m(\boldsymbol{X};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T}\middle|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}\right]=\mathbf{0}^T.$$

Because under suitable regularity, we have that $(\hat{\boldsymbol{\xi}}_n-\boldsymbol{\xi}^*)=O_p(n^{-1/2})$, it follows that the difference $n^{1/2}\{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)-\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n,\boldsymbol{\xi}^*)\}=o_p(1)$. Consequently, we find that

$$n^{1/2}\left\{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)-\mu_0\right\}=n^{-1/2}\sum_{i=1}^{n}\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O}_i)+o_p(1). \tag{2.34}$$

This means that under model $\mathcal{M}_{\text{SP}}$ (2.20), and thus that the model for the missingness mechanism is correctly specified, the estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)$ is RAL with influence function $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})$. This shows that the asymptotic behavior of this estimator is the same as if the unknown value $\boldsymbol{\xi}^*$ were known to us, regardless of correct specification of the working model $m(\boldsymbol{X};\boldsymbol{\xi})$, making the procedure adaptive.

The reason we use of the augmentation term $\{1-R/\pi(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n)\}m(\boldsymbol{X};\hat{\boldsymbol{\xi}}_n)$ is an attempt to improve the efficiency of the standard IPTW estimator $\hat{\mu}_{n,\text{IPTW}}$ by using the incomplete data; that is, using information from those individuals with $R_i=0$. The greatest gain in efficiency is obtained when $m(\boldsymbol{X};\hat{\boldsymbol{\xi}}_n)$ is a consistent estimator for the true conditional expectation $E(Y|\boldsymbol{X})$. In this case, $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})=\phi_{\text{eff}}(\boldsymbol{O})$, because the projection of the estimating function evaluated at the limits $\boldsymbol{\psi}_0$ and $\boldsymbol{\xi}^*(=\boldsymbol{\xi}_0)$ onto $\Lambda_{\boldsymbol{\psi}}$ then equals zero. Thus, when $m(\boldsymbol{X};\boldsymbol{\xi}^*)$ equals $E(Y|\boldsymbol{X})$, we can act as if the unknown value $\boldsymbol{\psi}_0$ were known to us. We furthermore have that in this case the estimator attains the semiparametric efficiency bound. Moreover, we may expect that the better $m(\boldsymbol{X};\boldsymbol{\xi}^*)$ approximates the true conditional expectation $E(Y|\boldsymbol{X})$, the greater the gain in efficiency is. Note that consistency is not altered under misspecification of the model for the conditional mean outcome. We conclude with the following result.

Proposition 2.7 (Locally Efficient Estimator)**.** *Under model $\mathcal{M}_{SP}$ (2.20), the estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)$ given in (2.33) with $\hat{\boldsymbol{\psi}}_n$ the MLE of $\boldsymbol{\psi}_0$, has influence*

*function $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})$ given in Proposition 2.5. When $m(\boldsymbol{X};\boldsymbol{\xi}^*)$ is correctly specified and thus equals $E(Y|\boldsymbol{X})$, the influence function equals the efficient influence function $\phi_{eff}(\boldsymbol{O})$ given in Proposition 2.6. The estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n)$ is thus **locally efficient**; it has smallest asymptotic variance within the class of all estimators that are consistent and asymptotically normal under model $\mathcal{M}_{SP}$ (2.20), provided that $m(\boldsymbol{X};\boldsymbol{\xi}^*)$ is also correctly specified. The semiparametric efficiency bound is thus attained locally.*

Estimating the asymptotic variance

To end this chapter, we show how the influence function of the locally efficient estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n)$ can be used to estimate the asymptotic variance of the estimator via the so-called **sandwich estimator** (Stefanski and Boos 2002).

Because the estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n)$ admits the expansion (2.34), it is RAL and its asymptotic distribution is given by $N(0,\sigma^2)$ with asymptotic variance $\sigma^2 = E\{\phi^2_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})\}$, the variance of the influence function $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})$. This can be estimated using the sandwich estimator:

$$\hat{\sigma}^2_n = n^{-1}\sum_{i=1}^{n} \hat{\phi}^2_n(\boldsymbol{O}_i; \hat{\mu}_n, \hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n), \tag{2.35}$$

where

$$\begin{aligned}
\hat{\phi}_n(\boldsymbol{O}_i; \hat{\mu}_n, \hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n) &= U(\boldsymbol{O}_i; \hat{\mu}_n, \hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n) \\
&\qquad - \hat{E}_n(U\boldsymbol{S}^T_{\boldsymbol{\psi}}) \left\{\hat{E}_n(\boldsymbol{S}_{\boldsymbol{\psi}}\boldsymbol{S}^T_{\boldsymbol{\psi}})\right\}^{-1} \boldsymbol{S}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n), \\
U(\boldsymbol{O}_i; \hat{\mu}_n, \hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n) &= \frac{R_i Y_i}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)} + \left\{1 - \frac{R_i}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)}\right\} m(\boldsymbol{X}_i; \hat{\tilde{\boldsymbol{\xi}}}_n) - \hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n), \\
\hat{E}_n(U\boldsymbol{S}^T_{\boldsymbol{\psi}}) &= n^{-1}\sum_{i=1}^{n} U(\boldsymbol{O}_i; \hat{\mu}_n, \hat{\boldsymbol{\psi}}_n, \hat{\tilde{\boldsymbol{\xi}}}_n)\boldsymbol{S}^T_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n), \\
\hat{E}_n(\boldsymbol{S}_{\boldsymbol{\psi}}\boldsymbol{S}^T_{\boldsymbol{\psi}}) &= n^{-1}\sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n)\boldsymbol{S}^T_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n), \\
\boldsymbol{S}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n) &= \frac{R_i - \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)\{1 - \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)\}} \pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n),
\end{aligned}$$

with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n) = \partial\pi(\boldsymbol{X}_i;\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_n}$.

Inference

Given the estimator $\hat{\sigma}_n^2$ for the asymptotic variance of the estimator $\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\hat{\boldsymbol{\xi}}}_n)$ (that is, the estimated variance of the influence function $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})$), an asymptotic $(1-\alpha)100\%$ confidence interval (CI) and p-value can be calculated based on the asymptotic normality of the estimator. A $(1-\alpha)100\%$ CI is given by

$$\left[\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\hat{\boldsymbol{\xi}}}_n) \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}\right]$$

where $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1-\alpha/2$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. A p-value for the hypothesis test $H_0 : \mu = \tilde{\mu}$ versus $H_a : \mu \neq \tilde{\mu}$ for some $\tilde{\mu} \in \mathbb{R}$ can be calculated as

$$p = 2\left\{1 - \Phi\left(\left|\frac{\hat{\mu}_n(\hat{\boldsymbol{\psi}}_n, \hat{\hat{\boldsymbol{\xi}}}_n) - \tilde{\mu}}{\hat{\sigma}_n/\sqrt{n}}\right|\right)\right\}.$$

CHAPTER 3

Doubly Robust Estimation

3.1 Introduction

In the previous chapter, we reviewed the basic concepts concerning the geometry of semiparametric models and the semiparametric efficiency theory. This understanding was sufficient to identify the efficient influence function for the estimation of a population mean outcome $E(Y) = \mu_0$ with incomplete data, explainable by a set of measured auxiliary covariates (so that the MAR assumption holds), under a semiparametric model that assumes the true missingness mechanism is known to belong to a class of known functions, parameterized by a finite-dimensional parameter. Motivated by the shape of this efficient influence function, we constructed an algorithm to obtain a locally efficient estimator (2.33) of μ_0. This resulting estimator is not only locally efficient, but also possesses a remarkable and very attractive property: **double robustness**, on which we will elaborate in this chapter. This chapter is based on Vermeulen and Vansteelandt (2015a).

3.2 Nuisance Working Models

Estimation of many statistical parameters requires postulation of so-called **nuisance working models**: models not of primary scientific interest, but needed to obtain a well-behaved estimator of the target parameter in small to moderate sample sizes. For instance, in Section 2.5, we studied the problem where the outcome data are incomplete in a way that is explainable by measured covariates. Robins et al. (1994) showed that consistent estimation of this mean outcome μ_0 requires specification of at least one of the two following working models.

The first is a working model for the probability of observing the data, the missingness model – referred to as the **propensity score** throughout:

$$P(R = 1|\boldsymbol{X}) = \pi_0(\boldsymbol{X}) = \pi(\boldsymbol{X}; \boldsymbol{\psi}_0).$$

We assume that $\pi_0(\boldsymbol{X}) \geq \delta > 0$ with probability one (cf., positivity, van der Laan and Rose (2011), chap. 10), and that $\pi(\boldsymbol{X}; \boldsymbol{\psi})$ is a known function, smooth in $\boldsymbol{\psi}$, and $\boldsymbol{\psi}_0$ is an unknown s-dimensional parameter; e.g., a logistic regression model $\pi(\boldsymbol{X}; \boldsymbol{\psi}) = \text{expit}\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\}$ with $\boldsymbol{l}(\boldsymbol{X}) = (1, X_1, \ldots, X_p)^T$ ($s = p+1$), can be used. The model $\mathcal{M}(\boldsymbol{\psi})$ denotes the statistical model for the joint distribution $f_{\boldsymbol{O}}(\boldsymbol{o}) = f_{RY,R,\boldsymbol{X}}(ry, r, \boldsymbol{x})$ (2.19) of the observed data $\boldsymbol{O}$ induced by the working model $\{\pi(\boldsymbol{X}; \boldsymbol{\psi}) : \boldsymbol{\psi} \in \mathbb{R}^s\}$ for the propensity score. It is correctly specified when $f_{\boldsymbol{O},0}(\boldsymbol{o}) \in \mathcal{M}(\boldsymbol{\psi})$. Note that $\mathcal{M}(\boldsymbol{\psi})$ corresponds to $\mathcal{M}_{\text{SP}}$ from Section 2.5. Let $\hat{\boldsymbol{\psi}}_n$ denote an arbitrary root-n consistent and asymptotically normal estimator of the nuisance parameter $\boldsymbol{\psi}$, which can be for instance the MLE, solving (2.30). When $\mathcal{M}(\boldsymbol{\psi})$ is correctly specified, the inverse probability of treatment weighted (IPTW) estimator (Horvitz and Thompson 1952)

$$\hat{\mu}_{n,\text{IPTW}} = n^{-1} \sum_{i=1}^{n} \frac{R_i Y_i}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)} \tag{3.1}$$

is consistent for μ_0.

The second is a working model for the conditional mean outcome

$$E(Y|\boldsymbol{X}) = m_0(\boldsymbol{X}) = m(\boldsymbol{X}; \boldsymbol{\xi}_0),$$

where $m(\boldsymbol{X};\boldsymbol{\xi})$ is a known function, smooth in $\boldsymbol{\xi}$, and where $\boldsymbol{\xi}_0$ is an unknown r-dimensional parameter; e.g., a linear model $m(\boldsymbol{X};\boldsymbol{\xi}) = \boldsymbol{\xi}^T \boldsymbol{k}(\boldsymbol{X})$ with $\boldsymbol{k}(\boldsymbol{X}) = (1, X_1, \ldots, X_p)^T$ $(r = p+1)$ for a continuous outcome Y can be used. The model $\mathcal{M}(\boldsymbol{\xi})$ denotes the statistical model for the joint density $f_{\boldsymbol{O}}(\boldsymbol{o}) = f_{RY,R,\boldsymbol{X}}(ry, r, \boldsymbol{x})$ (2.19) of the observed data $\boldsymbol{O}$ induced by the working model $\{m(\boldsymbol{X};\boldsymbol{\xi}) : \boldsymbol{\xi} \in \mathbb{R}^r\}$ for the conditional mean outcome. It is correctly specified when $f_{\boldsymbol{O},0}(\boldsymbol{o}) \in \mathcal{M}(\boldsymbol{\xi})$. Let $\hat{\boldsymbol{\xi}}_n$ denote an arbitrary root-n consistent and asymptotically normal estimator of the nuisance parameter $\boldsymbol{\xi}$, which can be for instance the MLE, solving (2.32). When $\mathcal{M}(\boldsymbol{\xi})$ is correctly specified, the imputation (IMP) estimator

$$\hat{\mu}_{n,\text{IMP}} = n^{-1} \sum_{i=1}^{n} m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n) \tag{3.2}$$

is consistent for μ_0. Note that $\hat{\mu}_{n,\text{IMP}}$ involves extrapolation if the distributions of the auxiliary covariates $\boldsymbol{X}$ conditional on $R = 1$ and $R = 0$ differ.

3.3 Doubly Robust Estimation

A prevailing concern now is that misspecification of these nuisance working models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\xi})$ may induce bias in the estimator of the target parameter μ_0 (Robins 1999a). In many missing data and causal inference problems however, this concern of bias due to model misspecification can be lessened via the use of so-called **doubly robust estimators**. These consistently estimate the target parameter when at least one of two nuisance working models is correctly specified, regardless of which (Robins and Rotnitzky 2001). For this specific missing data problem, Scharfstein et al. (1999a) showed that a doubly robust (DR) estimator of μ_0 can be obtained as

$$\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) = n^{-1} \sum_{i=1}^{n} \left[\frac{R_i Y_i}{\pi(\boldsymbol{X}_i, \hat{\boldsymbol{\psi}}_n)} + \left\{ 1 - \frac{R_i}{\pi(\boldsymbol{X}_i, \hat{\boldsymbol{\psi}}_n)} \right\} m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n) \right] \tag{3.3}$$

$$= n^{-1} \sum_{i=1}^{n} \left[m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n) + \frac{R_i}{\pi(\boldsymbol{X}_i, \hat{\boldsymbol{\psi}}_n)} \left\{ Y_i - m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n) \right\} \right]. \tag{3.4}$$

The double robustness states that this estimator is consistent for μ_0 under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\xi})$: as soon as one, but not necessarily both working models are correctly specified, which we will demonstrate next. If $\mathcal{M}(\boldsymbol{\psi})$ holds, $\hat{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}_0$, the truth, in which case $\pi(\boldsymbol{X}; \hat{\boldsymbol{\psi}}_n) \xrightarrow{p} \pi(\boldsymbol{X}; \boldsymbol{\psi}_0) = \pi_0(\boldsymbol{X})$. However, if $\mathcal{M}(\boldsymbol{\psi})$ is misspecified, $\hat{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}^*$, for some constant $\boldsymbol{\psi}^*$, in which case $\pi(\boldsymbol{X}; \hat{\boldsymbol{\psi}}_n) \xrightarrow{p} \pi(\boldsymbol{X}; \boldsymbol{\psi}^*) \neq \pi_0(\boldsymbol{X})$. Likewise, if $\mathcal{M}(\boldsymbol{\xi})$ is correctly specified, $\hat{\boldsymbol{\xi}}_n \xrightarrow{p} \boldsymbol{\xi}_0$, the truth, and then $m(\boldsymbol{X}; \hat{\boldsymbol{\xi}}_n) \xrightarrow{p} m(\boldsymbol{X}; \boldsymbol{\xi}_0) = m_0(\boldsymbol{X})$, whereas, if $\mathcal{M}(\boldsymbol{\xi})$ is misspecified, $\hat{\boldsymbol{\xi}}_n \xrightarrow{p} \boldsymbol{\xi}^*$, for some constant $\boldsymbol{\xi}^*$, in which case $m(\boldsymbol{X}; \hat{\boldsymbol{\xi}}_n) \xrightarrow{p} m(\boldsymbol{X}; \boldsymbol{\xi}^*) \neq m_0(\boldsymbol{X})$.

(a) **Suppose $\mathcal{M}(\boldsymbol{\psi})$ is correctly specified.** From (3.3) and the uniform WLLN, it follows that the estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ converges in probability to

$$\begin{aligned} E\left[\frac{RY}{\pi_0(\boldsymbol{X})} + \left\{1 - \frac{R}{\pi_0(\boldsymbol{X})}\right\} m(\boldsymbol{X}; \boldsymbol{\xi}^*)\right] \\ = E\left\{\frac{YE(R|Y,\boldsymbol{X})}{\pi_0(\boldsymbol{X})}\right\} + E\left[\left\{1 - \frac{E(R|\boldsymbol{X})}{\pi_0(\boldsymbol{X})}\right\} m(\boldsymbol{X}; \boldsymbol{\xi}^*)\right] \\ = E\left\{Y\frac{E(R|\boldsymbol{X})}{\pi_0(\boldsymbol{X})}\right\} = \mu_0, \end{aligned}$$

where we used the MAR assumption and the fact that $\pi_0(\boldsymbol{X}) = E(R|\boldsymbol{X})$. Thus, under $\mathcal{M}(\boldsymbol{\psi})$, we obtain $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) \xrightarrow{p} \mu_0$.

(b) **Suppose $\mathcal{M}(\boldsymbol{\xi})$ is correctly specified.** From (3.4) and the uniform WLLN, it follows that the estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ converges in probability to

$$\begin{aligned} E\left[m_0(\boldsymbol{X}) + \frac{R}{\pi(\boldsymbol{X}; \boldsymbol{\xi}^*)}\{Y - m_0(\boldsymbol{X})\}\right] \\ = E(Y) + E\left[\frac{R}{\pi(\boldsymbol{X}; \boldsymbol{\xi}^*)}\{E(Y|R,\boldsymbol{X}) - m_0(\boldsymbol{X})\}\right] \\ = \mu_0 + E\left[\frac{R}{\pi(\boldsymbol{X}; \boldsymbol{\xi}^*)}\{E(Y|\boldsymbol{X}) - m_0(\boldsymbol{X})\}\right] = \mu_0, \end{aligned}$$

where we used the MAR assumption and the fact that $m_0(\boldsymbol{X}) = E(Y|\boldsymbol{X})$. Thus, under $\mathcal{M}(\boldsymbol{\xi})$, we obtain $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) \xrightarrow{p} \mu_0$.

The reliance on multiple nuisance working models, of which only one must be correctly specified, makes the doubly robust estimator a potential 'compromise'

estimator amidst competing estimators that each rely on a single, but different working model. It is seen from (3.3) and (3.4) that the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ forms a compromise between the IPTW estimator $\hat{\mu}_{n,\mathrm{IPTW}}$, that relies solely on a model for the probability of missingness, and the imputation-based estimator $\hat{\mu}_{n,\mathrm{IMP}}$, that relies on an imputation model for the incomplete outcome; arguably, the doubly robust estimator may therefore define the preferred analysis.

The appeal of doubly robust estimators surpasses the defining property of double protection against model misspecification. Many doubly robust estimators are locally efficient within a broad class of estimators. Indeed, observe that the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ precisely equals the locally efficient estimator (2.33) constructed in Section 2.5. Thus, if the intersection model $\mathcal{M}(\boldsymbol{\psi}) \cap \mathcal{M}(\boldsymbol{\xi})$ holds, i.e., both working models are correctly specified, the doubly robust estimator (3.3) is locally efficient under model $\mathcal{M}(\boldsymbol{\psi})$: it then has the smallest asymptotic variance within the class of all RAL estimators that are consistent and asymptotically normal under $\mathcal{M}(\boldsymbol{\psi})$, provided that also $\mathcal{M}(\boldsymbol{\xi})$ is correctly specified. Because of this, the use of doubly robust estimators has also been advocated in randomized trial analyses: by exploiting the known randomization probabilities, they make it possible to increase power via covariate adjustment without risking bias due to model misspecification (Tsiatis et al. 2008; Zhang et al. 2008; Moore and van der Laan 2009; Vermeulen et al. 2015).

The estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ is just one example of a doubly robust estimator. Since the seminal work by Scharfstein et al. (1999a) and Robins and Rotnitzky (2001), doubly robust estimators have been developed for a variety of statistical parameters. Bang and Robins (2005) give an overview of work on doubly robust estimation of the parameters indexing conditional mean models when the outcome data are incomplete, and of marginal treatment effects in causal inference models. In the missing data literature, doubly robust estimators have also been developed for, for instance, the mean of K-sample U-statistics (Schisterman and Rotnitzky 2001), for the estimation of the area under the receiver operating characteristic curve (Rotnitzky et al. 2006), for nonparametric regression (Wang et al. 2010) and for incomplete covariate problems (Tchetgen Tchetgen and Rotnitzky 2011). In the causal inference literature, doubly robust estimators have also been proposed for statistical interaction parameters (Vansteelandt et al. 2008), for controlled direct

effects (Goetgeluk et al. 2008) and natural direct and indirect effects in mediation analysis (Tchetgen Tchetgen and Shpitser 2012; Zheng and van der Laan 2012), for average effects of time varying treatments (Murphy et al. 2001), for optimal treatment regimes (Orellana et al. 2010), for attributable fractions (Sjölander and Vansteelandt 2011), for instrumental variables analysis (Okui et al. 2012), for parameters indexing proportional hazards models (Hyun et al. 2012), for quantile-based treatment effects (Zhang et al. 2012), and for the marginal probabilistic index (MPI) (see Chapter 7). Recently, they are also being considered by companies, such as Google and Microsoft, for policy optimization and evaluation in the context of content recommendation and internet advertising (Dudík et al. 2015).

3.4 Asymptotic Distribution of the Doubly Robust Estimator Under the Union Model

In Section 2.5.3, we derived the influence of $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ under correct specification of model $\mathcal{M}(\boldsymbol{\psi})$ and when $\hat{\boldsymbol{\psi}}_n$ was estimated via MLE. We found that under these conditions, the influence function was given by $\phi_{m(\boldsymbol{X};\boldsymbol{\xi}^*)}(\boldsymbol{O})$ (see (2.25) and (2.26)). The projection term (2.26) appropriately accounts for the estimation of the propensity score, provided $\hat{\boldsymbol{\psi}}_n$ constitutes the MLE of $\boldsymbol{\psi}$. Furthermore, under model $\mathcal{M}(\boldsymbol{\psi})$, the asymptotic behavior of the doubly robust estimator did not depend on the choice of the root-n consistent estimator $\hat{\boldsymbol{\xi}}_n$ of $\boldsymbol{\xi}$. This follows from the fact that $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \boldsymbol{\xi})$ is a consistent estimator of μ_0 under $\mathcal{M}(\boldsymbol{\psi})$, regardless the value of $\boldsymbol{\xi}$. We also found that this influence function reduces to the efficient influence function $\phi_{\text{eff}}(\boldsymbol{O})$ at the intersection model $\mathcal{M}(\boldsymbol{\psi}) \cap \mathcal{M}(\boldsymbol{\xi})$. In this case, the asymptotic behavior of the doubly robust estimator does not depend on the choice of both the root-n consistent estimators $\hat{\boldsymbol{\psi}}_n$ of $\boldsymbol{\psi}$ and $\hat{\boldsymbol{\xi}}_n$ of $\boldsymbol{\psi}$. Consequently, $\hat{\boldsymbol{\psi}}_n$ need not equal the MLE of $\boldsymbol{\psi}$.

The aforementioned properties are no coincidence and more generally follow from the double robustness of the estimator $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$. Below, we study the first-order asymptotic behavior of this doubly robust estimator under potential misspecification of both working models and identify its influence function under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\xi})$. The proposition below, which follows from stan-

dard results on M-estimation and is proved in Appendix 3.A, gives the asymptotic distribution of the doubly robust estimator (3.3) in the special case where $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\xi}}_n$ are solutions to estimating equations

$$\begin{aligned}\sum_{i=1}^{n} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\hat{\boldsymbol{\psi}}_n) &= \mathbf{0},\\ \sum_{i=1}^{n} \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}_i;\hat{\boldsymbol{\xi}}_n) &= \mathbf{0}.\end{aligned}$$

Choosing the estimating functions $\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}) = \boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi})$ and $\boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\xi}) = R\{Y - \boldsymbol{\xi}^T\boldsymbol{k}(\boldsymbol{X})\}\boldsymbol{k}(\boldsymbol{X})$ (when $m(\boldsymbol{X};\boldsymbol{\xi}) = \boldsymbol{\xi}^T\boldsymbol{X}$) yields the MLE of $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$. The more general result when the estimating functions $\boldsymbol{U}_{\boldsymbol{\psi}}$ and $\boldsymbol{U}_{\boldsymbol{\xi}}$ involve both $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ is reported in Appendix 3.A.

Proposition 3.1. *Define*

$$\phi(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi}) = \frac{RY}{\pi(\boldsymbol{X};\boldsymbol{\psi})} + \left\{1 - \frac{R}{\pi(\boldsymbol{X};\boldsymbol{\psi})}\right\} m(\boldsymbol{X},\boldsymbol{\xi}) - \mu \qquad (3.5)$$

and denote $\boldsymbol{\psi}^* = plim(\hat{\boldsymbol{\psi}}_n)$, $\boldsymbol{\xi}^* = plim(\hat{\boldsymbol{\xi}}_n)$, i.e., the probability limits of estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\xi}}_n$ of $\boldsymbol{\psi}_0$ and $\boldsymbol{\xi}_0$; these probability limits may differ from the truth $\boldsymbol{\psi}_0$ and $\boldsymbol{\xi}_0$ when the working models are misspecified, but not otherwise. Under suitable regularity conditions (Robins et al. 1994, app. B), $\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)$ is asymptotically linear with influence function

$$\begin{aligned}&\tilde{\phi}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\\ &\quad= \phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\\ &\qquad - E\left\{\frac{\partial\phi}{\partial\boldsymbol{\psi}^T}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\right\} E^{-1}\left\{\frac{\partial\boldsymbol{U}_{\boldsymbol{\psi}}}{\partial\boldsymbol{\psi}^T}(\boldsymbol{O};\boldsymbol{\psi}^*)\right\}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)\\ &\qquad - E\left\{\frac{\partial\phi}{\partial\boldsymbol{\xi}^T}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\right\} E^{-1}\left\{\frac{\partial\boldsymbol{U}_{\boldsymbol{\xi}}}{\partial\boldsymbol{\xi}^T}(\boldsymbol{O};\boldsymbol{\xi}^*)\right\}\boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\xi}^*);\end{aligned}$$

meaning $\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ *can be expanded as*

$$n^{1/2}\{\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \mu_0\} = n^{-1/2}\sum_{i=1}^{n} \tilde{\phi}(\boldsymbol{O}_i; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) + o_p(1),$$

where $o_p(1)$ *denotes a term that converges to zero in probability. Consequently,*

$$n^{1/2}[\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \mu_0 - E\{\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)\}]$$

converges to a normal limit with mean zero and variance $var(\tilde{\phi})$.

It is interesting to investigate the mean gradients $E\{\partial\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)/\partial\boldsymbol{\psi}^T\}$ and $E\{\partial\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)/\partial\boldsymbol{\xi}^T\}$ to gain a better understanding in how working model misspecification affects the first-order asymptotic behavior of the doubly robust estimator. With $\pi_{\boldsymbol{\psi}}(\boldsymbol{X}; \boldsymbol{\psi}^*) = \partial\pi(\boldsymbol{X}; \boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}$ and $m_{\boldsymbol{\xi}}(\boldsymbol{X}; \boldsymbol{\xi}^*) = \partial m(\boldsymbol{X}; \boldsymbol{\xi})/\partial\boldsymbol{\xi}|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}$, we obtain from

$$\begin{aligned}
E\left\{\frac{\partial\phi}{\partial\boldsymbol{\psi}^T}(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)\right\} &= -E\left[\frac{R}{\pi^2(\boldsymbol{X}; \boldsymbol{\psi}^*)}\{Y - m(\boldsymbol{X}; \boldsymbol{\xi}^*)\}\pi^T_{\boldsymbol{\psi}}(\boldsymbol{X}; \boldsymbol{\psi}^*)\right] \\
&= E\left[\frac{\pi_0(\boldsymbol{X})}{\pi^2(\boldsymbol{X}; \boldsymbol{\psi}^*)}\left\{m(\boldsymbol{X}; \boldsymbol{\xi}^*) - m_0(\boldsymbol{X})\right\}\pi^T_{\boldsymbol{\psi}}(\boldsymbol{X}; \boldsymbol{\psi}^*)\right], \\
E\left\{\frac{\partial\phi}{\partial\boldsymbol{\xi}^T}(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)\right\} &= E\left[\left\{1 - \frac{R}{\pi(\boldsymbol{X}; \boldsymbol{\psi}^*)}\right\}m^T_{\boldsymbol{\xi}}(\boldsymbol{X}; \boldsymbol{\xi}^*)\right] \\
&= E\left[\left\{1 - \frac{\pi_0(\boldsymbol{X})}{\pi(\boldsymbol{X}; \boldsymbol{\psi}^*)}\right\}m^T_{\boldsymbol{\xi}}(\boldsymbol{X}; \boldsymbol{\xi}^*)\right],
\end{aligned}$$

that $E\{\partial\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)/\partial\boldsymbol{\psi}^T\} = \mathbf{0}$ and $E\{\partial\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)/\partial\boldsymbol{\xi}^T\} = \mathbf{0}$ at the intersection model $\mathcal{M}(\boldsymbol{\psi}) \cap \mathcal{M}(\boldsymbol{\xi})$ since then for both working models $\pi_0(\boldsymbol{X}) = \pi(\boldsymbol{X}; \boldsymbol{\psi}^*)$ and $m_0(\boldsymbol{X}) = m(\boldsymbol{X}; \boldsymbol{\xi}^*)$. The influence function of the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ then simply becomes $\tilde{\phi}(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}_0, \boldsymbol{\xi}_0) = \phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}_0, \boldsymbol{\xi}_0)$, where $\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}, \boldsymbol{\xi})$ is the influence function of $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi}, \boldsymbol{\xi})$; i.e., the doubly robust estimator of μ_0 evaluated at fixed nuisance parameter values $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$. Note that, coinciding with the results from Section 2.5, this influence function equals the efficient influence function $\phi_{\mathrm{eff}}(\boldsymbol{O})$ (2.28). We can conclude that, under correctly specified working models, the choice of root-n consistent estimators of the nuisance

parameters does not affect the first-order asymptotic distribution of the doubly robust estimator. This property, which is more generally satisfied for doubly robust estimators (Robins and Rotnitzky 2001), has stimulated the use of standard methods, such as maximum likelihood estimation, to estimate the nuisance parameters (Bang and Robins 2005).

Under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\xi})$ however, only one of both mean gradients is zero. For instance, suppose that $\mathcal{M}(\boldsymbol{\psi})$ holds, but $\mathcal{M}(\boldsymbol{\xi})$ can be misspecified. In this case, it is guaranteed that $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)/\partial\boldsymbol{\xi}^T\} = \mathbf{0}$, and hence, the choice of the root-n consistent estimator of $\boldsymbol{\xi}$ does not affect the first-order asymptotic behavior of the doubly robust estimator (see also Section 2.5.3). Because $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)/\partial\boldsymbol{\psi}^T\} \neq \mathbf{0}$, the choice of the root-n consistent estimator $\hat{\boldsymbol{\psi}}_n$ of $\boldsymbol{\psi}$ does affect the first-order asymptotic behavior of $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)$. E.g., when $\hat{\boldsymbol{\psi}}_n$ is the MLE of $\boldsymbol{\psi}$, $\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}) = \boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi})$ and the correction term for the estimation of $\boldsymbol{\psi}$ equals the orthogonal projection of $\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)$ onto the nuisance tangent space $\Lambda_{\boldsymbol{\psi}}$: because

$$
\begin{aligned}
E\left\{\frac{\partial\phi}{\partial\boldsymbol{\psi}^T}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)\right\} &= -E\left\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right\},\\
E\left\{\frac{\partial\boldsymbol{S}_{\boldsymbol{\psi}}}{\partial\boldsymbol{\psi}^T}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right\} &= -E\left\{\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right\},
\end{aligned}
$$

with $E\left\{\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right\} = I_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0)$ denoting the Fisher information matrix for $\boldsymbol{\psi}$, we get

$$
\begin{aligned}
&-E\left\{\frac{\partial\phi}{\partial\boldsymbol{\psi}^T}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)\right\}E^{-1}\left\{\frac{\partial\boldsymbol{S}_{\boldsymbol{\psi}}}{\partial\boldsymbol{\psi}^T}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right\}\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\\
&\quad= -E\left\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)\boldsymbol{S}_{\boldsymbol{\psi}}^T(R,\boldsymbol{X};\boldsymbol{\psi}_0)\right\}I_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}_0)\boldsymbol{S}_{\boldsymbol{\psi}}(R,\boldsymbol{X};\boldsymbol{\psi}_0)\\
&\quad= -\Pi\left\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)\middle|\Lambda_{\boldsymbol{\psi}}\right\},
\end{aligned}
$$

so that Proposition 3.1 reduces to Proposition 2.7.

3.5 Discussion

Estimation of the nuisance parameters indexing the working models in doubly robust estimators has long been ignored. We observed in Section 3.4 that the choice of nuisance parameter estimators has no impact on the asymptotic variance of doubly robust estimators when both working models are correctly specified. This has lead to the default use of maximum likelihood estimators. This standard practice has gradually started to change when simulation studies by Kang and Schafer (2007a) cautioned for potentially disastrous performance of certain doubly robust estimators (relative to simpler estimators) when both working models are misspecified. The discussions of that article (Ridgeway and McCaffrey 2007; Robins et al. 2007; Tan 2007; Tsiatis and Davidian 2007; Kang and Schafer 2007b) reveal that many different doubly robust estimators may exist for a given target parameter, all with potentially very different behavior and properties under misspecification of at least one working model. In particular, when a doubly robust estimator exists for a given target parameter, then infinitely many can usually be constructed by varying the choice of nuisance parameter estimators (as shown in Section 3.4, Proposition 3.1). All resulting doubly robust estimators have the same asymptotic behavior under correct specification of all working models, but a potentially very different behavior under model misspecification.

Recently, many alternatives have been proposed in the literature. Rubin and van der Laan (2008), Cao et al. (2009) and Tsiatis et al. (2011) developed doubly robust estimators in specific missing data models with desirable efficiency properties when the missingness model is correctly specified. In their development, which generalizes that of Tan (2006), the nuisance parameters indexing the working model for the incomplete outcome are estimated by directly minimizing the variance of the doubly robust estimator. The TMLE (targeted maximum likelihood estimation and more general targeted minimum loss-based estimation) procedure (van der Laan and Rose 2011) and the procedures of Tan (2010) and Rotnitzky et al. (2012) guarantee doubly robust estimators that fall within the allowed parameter range, with the latter procedures also having desirable efficiency properties. With the exception of TMLE, all these proposals focus on improving the efficiency of doubly robust estimators under misspecification of the working model for the full-data distribution

(i.e., the dependence of outcome on covariates/confounders in missing data/causal inference models). The collaborative TMLE (C-TMLE) procedure of van der Laan and Gruber (2010), which is a further improvement upon the TMLE procedure, additionally focuses on the estimation of the missingness/exposure probabilities in a way that aims to prevent large variance of the doubly robust estimator. We will review some of these alternatives in Chapter 4 and for more details, we refer to the literature, e.g., Rotnitzky and Vansteelandt (2014).

In this thesis, as in a recent paper by van der Laan (2014), we take a different perspective by focusing on **bias reduction** rather than variance reduction. This is motivated by the fact that the finite-sample bias of a doubly robust estimator can get severely amplified under misspecification of at least one working model and may become especially severe under misspecification of both working models (Kang and Schafer 2007a; Vansteelandt et al. 2012). In particular, in the next chapter, we propose a general estimating equation based strategy, referred to as **bias-reduced doubly robust estimation**, which locally minimizes the squared first-order asymptotic bias of the doubly robust estimator in the direction of the nuisance parameters under misspecification of both working models. This proposal differs from van der Laan (2014) in that it avoids bias approximations in view of the difficulty of approximating bias. Unlike most other proposals, it focuses on the estimation of the nuisance parameters in **all** working models, is readily applicable to arbitrary doubly robust estimators, can be adapted to certain multiply robust estimators and returns doubly robust estimators with easy-to-calculate asymptotic variance.

3.A Asymptotic Normality of $\hat{\mu}_{n,\mathbf{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$

Define the estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\xi}}_n$ for the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ to be the solution to conformable estimating equations

$$\sum_{i=1}^{n} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) = \mathbf{0},$$
$$\sum_{i=1}^{n} \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) = \mathbf{0}$$

and let $\boldsymbol{\psi}^* = \text{plim}(\hat{\boldsymbol{\psi}}_n)$ and $\boldsymbol{\xi}^* = \text{plim}(\hat{\boldsymbol{\xi}}_n)$.

Proposition 3.2. *Under suitable regularity conditions (Robins et al. 1994, app. B), a first-order asymptotic expansion of the doubly robust estimator $\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ is given by*

$$n^{1/2}\{\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \mu_0\} = n^{-1/2} \sum_{i=1}^{n} \tilde{\phi}(\boldsymbol{O}_i; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) + o_p(1),$$

where

$$\begin{aligned}
&\tilde{\phi}(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \\
&= \phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \\
&- E\left(\frac{\partial \phi}{\partial \boldsymbol{\psi}^T}\right) \left\{ \mathbf{I}_s - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right) \right\}^{-1} \\
&\times \left\{ E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right. \\
&\qquad \left. - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right\} \\
&- E\left(\frac{\partial \phi}{\partial \boldsymbol{\xi}^T}\right) \left\{ \mathbf{I}_r - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) \right\}^{-1} \\
&\times \left\{ E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right.
\end{aligned}$$

$$-E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right)E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right)E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right)\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\Bigg\},$$

where all gradients are evaluated at $(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)$. *Consequently,*

$$n^{1/2}[\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n)-\mu_0-E\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\}] \xrightarrow{d} N\{0,var(\tilde{\phi})\},$$

where $\xrightarrow{d}$ *denotes convergence in distribution.*

Proof. Denote $\hat{\mu}_{n,\mathrm{DR}} \equiv \hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\hat{\boldsymbol{\xi}}}_n)$. From a standard Taylor expansion of $\phi(\boldsymbol{O};\hat{\mu}_{n,\mathrm{DR}},\hat{\boldsymbol{\psi}}_n,\hat{\hat{\boldsymbol{\xi}}}_n)$ around μ_0, $\boldsymbol{\psi}^*$ and $\boldsymbol{\xi}^*$ and the uniform WLLN (see Newey and McFadden (1994), Lemma 4.3), assuming suitable regularity conditions (Robins et al. 1994, app. B), we obtain

$$\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^{n}\phi(\boldsymbol{O}_i;\hat{\mu}_{n,\mathrm{DR}},\hat{\boldsymbol{\psi}}_n,\hat{\hat{\boldsymbol{\xi}}}_n)\\
&= n^{-1/2}\sum_{i=1}^{n}\phi\left(\boldsymbol{O}_i;\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*\right)+E\left(\frac{\partial\phi}{\partial\mu}\right)n^{1/2}(\hat{\mu}_{n,\mathrm{DR}}-\mu_0)\\
&\quad +E\left(\frac{\partial\phi}{\partial\boldsymbol{\psi}^T}\right)n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}^*)+E\left(\frac{\partial\phi}{\partial\boldsymbol{\xi}^T}\right)n^{1/2}(\hat{\hat{\boldsymbol{\xi}}}_n-\boldsymbol{\xi}^*)+o_p(1).
\end{aligned}$$

Consequently,

$$\begin{aligned}
n^{1/2}(\hat{\mu}_{n,\mathrm{DR}}-\mu_0) = n^{-1/2}\sum_{i=1}^{n}&\phi(\boldsymbol{O}_i;\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\\
&+E\left(\frac{\partial\phi}{\partial\boldsymbol{\psi}^T}\right)n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}^*)\\
&+E\left(\frac{\partial\phi}{\partial\boldsymbol{\xi}^T}\right)n^{1/2}(\hat{\hat{\boldsymbol{\xi}}}_n-\boldsymbol{\xi}^*)+o_p(1).
\end{aligned}$$

Next, using standard arguments and suitable regularity conditions,

$$n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}^*) = -n^{-1/2}E^{-1}\left(\frac{\partial\boldsymbol{U}_{\boldsymbol{\psi}}}{\partial\boldsymbol{\psi}^T}\right)\sum_{i=1}^{n}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)$$

$$- E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*) + o_p(1),$$

$$n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*) = -n^{-1/2} E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) \sum_{i=1}^{n} \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}_i; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)$$
$$- E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right) n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) + o_p(1).$$

Solving for $n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*)$ and $n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*)$, gives

$$\begin{aligned}
&n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) \\
&= -\left\{ \mathbf{I}_s - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right) \right\}^{-1} \\
&\quad \times n^{-1/2} \sum_{i=1}^{n} \left\{ E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right. \\
&\qquad \left. - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}_i; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right\} + o_p(1), \\
&n^{1/2}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*) \\
&= -\left\{ \mathbf{I}_r - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\xi}^T}\right) \right\}^{-1} \\
&\quad \times n^{-1/2} \sum_{i=1}^{n} \left\{ E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O}_i; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right. \\
&\qquad \left. - E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}\right) E\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\psi}^T}\right) E^{-1}\left(\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}\right) \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \boldsymbol{\psi}^*, \boldsymbol{\xi}^*) \right\} + o_p(1),
\end{aligned}$$

where $\mathbf{I}_s$ and $\mathbf{I}_r$ are identity matrices of dimension s and r, respectively. Substituting these asymptotic expansions in the expansion of $\hat{\mu}_{n,\mathrm{DR}}$, we conclude that the influence function of the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ equals $\tilde{\phi}(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)$. From the obtained asymptotic linearity of the doubly robust estimator and the CLT, it follows that $n^{1/2}[\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n) - \mu_0 - E\{\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)\}]$ converges to a normal limit with variance $\mathrm{var}(\tilde{\phi})$. □

When estimating equations of the type $\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi})$ and $\boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\xi})$ are used, as in Proposition 3.1, then

$$\begin{aligned} \partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi})/\partial \boldsymbol{\xi}^T &\equiv \mathbf{0}, \\ \partial \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\xi})/\partial \boldsymbol{\psi}^T &\equiv \mathbf{0}, \end{aligned}$$

and the influence function of $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\xi}}_n)$ reduces to

$$\begin{aligned} &\tilde{\phi}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*) \\ &\quad = \phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*) \\ &\qquad - E\left\{\frac{\partial \phi}{\partial \boldsymbol{\psi}^T}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\right\} E^{-1}\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}^T}(\boldsymbol{O};\boldsymbol{\psi}^*)\right\} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*) \\ &\qquad - E\left\{\frac{\partial \phi}{\partial \boldsymbol{\xi}^T}(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\right\} E^{-1}\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}^T}(\boldsymbol{O};\boldsymbol{\xi}^*)\right\} \boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\xi}^*). \end{aligned}$$

CHAPTER 4

Bias-Reduced Doubly Robust Estimation

4.1 Introduction

Over the past decade, doubly robust estimators have been proposed for a variety of target parameters in causal inference and missing data models. These are asymptotically unbiased when at least one of two nuisance working models is correctly specified, regardless of which (see Proposition 3.1 and Robins and Rotnitzky 2001). While their asymptotic distribution is not affected by the choice of root-n consistent estimators of the nuisance parameters indexing these working models when all working models are correctly specified, this choice of estimators can have a dramatic impact under misspecification of at least one working model.

In this chapter, we will propose a fairly simple and generic estimation principle for the nuisance parameters indexing each of the working models, which is designed to improve the performance of the doubly robust estimator of interest, relative to the default use of maximum likelihood estimators for the nuisance parameters. The proposed approach **locally minimizes the squared first-order asymptotic bias** of the doubly robust estimator under misspecification of both working models and results in doubly robust estimators with easy-to-calculate asymptotic variance. It

moreover improves the stability of the weights in those doubly robust estimators that invoke inverse probability weighting. Simulation studies confirm the desirable finite-sample performance of the proposed estimators. This chapter is based on Vermeulen and Vansteelandt (2015a).

4.2 Biased-Reduced Doubly Robust Estimation

The property that the choice of root-n consistent estimators of the nuisance parameters does not affect the first-order asymptotic behavior of a doubly robust estimator, is lost as soon as one of both working models is misspecified. Starting from a given doubly robust estimator, infinitely many doubly robust estimators can therefore typically be constructed by varying the choice of nuisance parameter estimators. This calls for estimation strategies for the nuisance parameters that are optimal according to some criterion. In this chapter, we propose nuisance parameter estimators whose probability limits locally minimize the squared first-order asymptotic bias of the doubly robust estimator under misspecification of both working models.

To make the presentation as general as possible (with a slight abuse of notation to simplify), let μ_0 denote the (unknown) population value of the scalar target parameter and $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\xi})$ a doubly robust estimator for it, based on finite-dimensional working models, inducing the semiparametric models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\xi})$ for the observed data distribution (see Section 3.2), indexed by parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ which take on the values $\boldsymbol{\psi}_0$ and $\boldsymbol{\xi}_0$ when, respectively, $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\xi})$ hold. The observed data is denoted $\{\boldsymbol{O}_i : i = 1,\ldots,n\}$. Finally, $\phi(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi})$ denotes the influence function of the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\xi})$ for fixed values of the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$.

4.2.1 Proposal

Consider possibly misspecified working models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\xi})$ at fixed known values $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, respectively. The first-order asymptotic bias of the doubly robust estimator is then given by $\mathrm{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0) = E\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})\}$. In the missing data problem of Section 3.3, this is also the exact finite-sample bias. By the double robustness, $\mathrm{bias}(\boldsymbol{\psi}_0,\boldsymbol{\xi};\mu_0) = 0$ for any $\boldsymbol{\xi}$ under $\mathcal{M}(\boldsymbol{\psi})$ and $\mathrm{bias}(\boldsymbol{\psi},\boldsymbol{\xi}_0;\mu_0) = 0$

for any $\boldsymbol{\psi}$ under $\mathcal{M}(\boldsymbol{\xi})$. This property is lost when both nuisance working models are misspecified. Suppose now that there exists a vector $(\boldsymbol{\psi}^{*,T}_{\text{BR}}, \boldsymbol{\xi}^{*,T}_{\text{BR}})^T$ such that $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{\text{BR}},\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\psi}\} = \mathbf{0}$ and $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{\text{BR}},\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\xi}\} = \mathbf{0}$, with BR an abbreviation for Bias-Reduced. The following theorem then shows that under suitable regularity conditions $(\boldsymbol{\psi}^{*,T}_{\text{BR}}, \boldsymbol{\xi}^{*,T}_{\text{BR}})^T$ locally minimizes the squared first-order bias in the direction of $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, explaining the subscript BR.

Theorem 4.1. *Under suitable regularity conditions (which are given in Appendix 4.A), $(\boldsymbol{\psi}^{*,T}_{BR}, \boldsymbol{\xi}^{*,T}_{BR})^T$ locally minimizes the squared first-order asymptotic bias*

$$bias^2(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0) = E^2\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})\},$$

with $(\boldsymbol{\psi}^{*,T}_{BR}, \boldsymbol{\xi}^{*,T}_{BR})^T$ a solution to $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{BR},\boldsymbol{\xi}^*_{BR})/\partial\boldsymbol{\psi}\} = \mathbf{0}$ and $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{BR},\boldsymbol{\xi}^*_{BR})/\partial\boldsymbol{\xi}\} = \mathbf{0}$.

Proof. Under regularity conditions that allow us to interchange integration and differentiation (see Appendix 4.A),

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\psi}}\text{bias}^2(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0) &= 2\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)\frac{\partial}{\partial\boldsymbol{\psi}}\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)\\
&= 2\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)E\left\{\frac{\partial\phi}{\partial\boldsymbol{\psi}}(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})\right\},\\
\frac{\partial}{\partial\boldsymbol{\xi}}\text{bias}^2(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0) &= 2\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)\frac{\partial}{\partial\boldsymbol{\xi}}\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)\\
&= 2\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)E\left\{\frac{\partial\phi}{\partial\boldsymbol{\xi}}(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})\right\}.
\end{aligned}$$

The result follows since by definition the values $\boldsymbol{\psi}^*_{\text{BR}}$ and $\boldsymbol{\xi}^*_{\text{BR}}$ satisfy the equations $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{\text{BR}},\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\psi}\} = \mathbf{0}$ and $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{\text{BR}},\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\xi}\} = \mathbf{0}$. □

In practice, the values $(\boldsymbol{\psi}^{*,T}_{\text{BR}}, \boldsymbol{\xi}^{*,T}_{\text{BR}})^T$ that solve the mean gradients

$$E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{\text{BR}},\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\psi}\} = \mathbf{0}, \tag{4.1}$$

$$E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*_{\text{BR}},\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\xi}\} = \mathbf{0} \tag{4.2}$$

are unknown and need to be estimated. Therefore, define the estimators $\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}$ as the solutions to the estimating equations

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\xi}} \phi(\boldsymbol{O}_i; \mu_0, \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}) = \mathbf{0}, \tag{4.3}$$

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\psi}} \phi(\boldsymbol{O}_i; \mu_0, \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}) = \mathbf{0}. \tag{4.4}$$

When the gradient of $\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}, \boldsymbol{\xi})$ with respect to $\boldsymbol{\psi}$ or $\boldsymbol{\xi}$ depends on the unknown population value μ_0, a preliminary consistent doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}^{\mathrm{prel}}$ is substituted for μ_0 (e.g., the default doubly robust estimator based on MLEs for the nuisance parameters).

The following theorem shows that (4.3) and (4.4) yield consistent estimators $\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}$ for $\boldsymbol{\psi}_0$ and $\boldsymbol{\xi}_0$ under models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\xi})$, respectively. The resulting doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ is referred to as the **bias-reduced doubly robust estimator**.

> **Theorem 4.2.** *Under suitable regularity conditions (see Appendix 4.A),* $\hat{\boldsymbol{\psi}}_n^{BR}$ *is a consistent estimator of* $\boldsymbol{\psi}_0$ *under* $\mathcal{M}(\boldsymbol{\psi})$ *and* $\hat{\boldsymbol{\xi}}_n^{BR}$ *is a consistent estimator of* $\boldsymbol{\xi}_0$ *under* $\mathcal{M}(\boldsymbol{\xi})$.

Proof. We give the proof for $\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}$. Under model $\mathcal{M}(\boldsymbol{\psi})$, the (unknown) population value $\boldsymbol{\psi}_0$ is well defined. By the double robustness of the estimator $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi}, \boldsymbol{\xi})$, $\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}_0, \boldsymbol{\xi})$ has mean zero for all $\boldsymbol{\xi}$. Consequently, with $F_0(\boldsymbol{o})$ denoting the true (unknown) joint distribution function of $\boldsymbol{O}$,

$$\begin{aligned} \mathbf{0} = \frac{\partial}{\partial \boldsymbol{\xi}} E\{\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}_0, \boldsymbol{\xi})\} &= \int \frac{\partial}{\partial \boldsymbol{\xi}} \phi(\boldsymbol{o}; \mu_0, \boldsymbol{\psi}_0, \boldsymbol{\xi}) \mathrm{d}F_0(\boldsymbol{o}) \\ &= E\left\{ \frac{\partial}{\partial \boldsymbol{\xi}} \phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}_0, \boldsymbol{\xi}) \right\} \end{aligned}$$

for all $\boldsymbol{\xi}$ assuming we can interchange integration and differentiation (see Appendix 4.A). Note that we do not take derivatives of $F_0(\boldsymbol{O})$ with respect to $\boldsymbol{\xi}$, since the expectation is taken under the true data generating mechanism, which stays fixed as $\boldsymbol{\xi}$ varies. Hence, at $\mathcal{M}(\boldsymbol{\psi})$, the gradient $\partial \phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}, \boldsymbol{\xi}) / \partial \boldsymbol{\xi}$ yields an unbiased

estimating function for $\boldsymbol{\psi}$. Under suitable regularity conditions (Robins et al. 1994, app. B), it now follows from the uniform WLLN (Newey and McFadden 1994, lemma 4.3) and the fact that $\text{plim}(\hat{\mu}^{\text{prel}}_{n,\text{DR}}) = \mu_0$ under $\mathcal{M}(\boldsymbol{\psi})$ that

$$\mathbf{0} = \text{plim}\left\{ n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\xi}} \phi(\boldsymbol{O}_i; \hat{\mu}^{\text{prel}}_{n,\text{DR}}, \hat{\boldsymbol{\psi}}^{\text{BR}}_n, \hat{\boldsymbol{\xi}}^{\text{BR}}_n) \right\} = E\left\{ \frac{\partial}{\partial \boldsymbol{\xi}} \phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}^*_{\text{BR}}, \boldsymbol{\xi}^*_{\text{BR}}) \right\}$$

with probability limits $\boldsymbol{\psi}^*_{\text{BR}} = \text{plim}(\hat{\boldsymbol{\psi}}^{\text{BR}}_n)$ and $\boldsymbol{\xi}^*_{\text{BR}} = \text{plim}(\hat{\boldsymbol{\xi}}^{\text{BR}}_n)$ so that $\boldsymbol{\psi}_0 = \boldsymbol{\psi}^*_{\text{BR}}$ and thus $\hat{\boldsymbol{\psi}}^{\text{BR}}_n$ is a consistent estimator of $\boldsymbol{\psi}_0$. The proof of the consistency of $\hat{\boldsymbol{\xi}}^{\text{BR}}_n$ is analogous. □

Theorem 4.2 shows that $\boldsymbol{\psi}^*_{\text{BR}} = \boldsymbol{\psi}_0$ under $\mathcal{M}(\boldsymbol{\psi})$ and $\boldsymbol{\xi}^*_{\text{BR}} = \boldsymbol{\xi}_0$ under $\mathcal{M}(\boldsymbol{\xi})$. This is not surprising because $\min_{(\boldsymbol{\psi}^T, \boldsymbol{\xi}^T)^T} \left\{ \text{bias}^2(\boldsymbol{\psi}, \boldsymbol{\xi}; \mu_0) \right\} = 0$ under $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\xi})$. Furthermore, we also have that

$$\text{bias}(\hat{\boldsymbol{\psi}}^{\text{BR}}_n, \hat{\boldsymbol{\xi}}^{\text{BR}}_n; \mu_0) = \text{bias}(\boldsymbol{\psi}^*_{\text{BR}}, \boldsymbol{\xi}^*_{\text{BR}}; \mu_0) + o(1)$$

(see Appendix 4.B), and hence, the squared first-order asymptotic bias is also locally minimized when the fixed values $(\boldsymbol{\psi}^{*,T}_{\text{BR}}, \boldsymbol{\xi}^{*,T}_{\text{BR}})^T$ are replaced by root-n consistent estimators $(\hat{\boldsymbol{\psi}}^{\text{BR},T}_n, \hat{\boldsymbol{\xi}}^{\text{BR},T}_n)^T$. However, the bias optimality promised by Theorem 4.1 may become somewhat illusory when the estimating equations (4.3) or (4.4) depend on the population value μ_0. The reason is that in this case, the values $(\boldsymbol{\psi}^{*,T}_{\text{BR}}, \boldsymbol{\xi}^{*,T}_{\text{BR}})^T$ no longer minimize $\text{bias}^2(\boldsymbol{\psi}, \boldsymbol{\xi}; \mu_0) = E\{\phi(\boldsymbol{O}; \mu_0, \boldsymbol{\psi}, \boldsymbol{\xi})\}^2$ but instead minimize $\text{bias}^2(\boldsymbol{\psi}, \boldsymbol{\xi}; \mu^*) = E\{\phi(\boldsymbol{O}; \mu^*, \boldsymbol{\psi}, \boldsymbol{\xi})\}^2$ with $\mu^* = \text{plim}(\hat{\mu}^{\text{prel}}_{n,\text{DR}})$, which may differ from μ_0 under misspecification of both nuisance working models. The bias optimality of Theorem 4.1 is therefore limited to doubly robust estimators for which the left-hand side of (4.3) and the left-hand side (4.4) do not depend on the target parameter. Fortunately, many target parameters for which doubly robust estimators exist satisfy this property. For instance, the class given in Robins et al. (2008, sec. 3.1) satisfies this; in particular, this is satisfied for the missing data example in Section 3.3 (see equation (3.5)). When the left-hand side of (4.3) and the left-hand side (4.4) do depend on the target parameter, then the bias optimality of Theorem 4.1 remains useful for score tests of the null hypothesis that $\mu = \tilde{\mu}$ for

some $\tilde{\mu}$; i.e., tests of $E\{\phi(\boldsymbol{O};\tilde{\mu},\boldsymbol{\psi}_0,\boldsymbol{\xi}_0)\}=0$. When μ_0 is substituted by $\tilde{\mu}$ in (4.3) and (4.4), the resulting (probability limits of the) estimators $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{BR}}$ continue to minimize $E^2\{\phi(\boldsymbol{O};\tilde{\mu},\boldsymbol{\psi},\boldsymbol{\xi})\}$.

A limitation of the proposed approach is that it demands working models of the same dimension because the gradient of ϕ with respect to $\boldsymbol{\xi}$ is used as an estimating function for $\boldsymbol{\psi}$ and vice versa. Restrictions on the dimension of the working models are also seen in other proposals (Rotnitzky et al. 2012). This can be remedied by enlarging the working models with clever choices of covariates until they reach the same dimension (see Section 4.7). An alternative is to minimize the squared first-order asymptotic bias in the direction of a single nuisance parameter, for instance in the dimension of $\boldsymbol{\psi}$ (see Chapter 5). Do note that in some cases, such as for the missing data problem introduced in Section 3.3, this restriction may not be that severe in practice. This is because many practitioners would anyway use models with the same covariates and main effects only.

4

Remark 4.1. *The validity of the proposal is predicated on the availability of a doubly robust estimator; it cannot be used for arbitrary estimators. Indeed, reconsider the missing data problem of Section 3.2 with* $\hat{\mu}_{n,IMP}(\boldsymbol{\xi})=n^{-1}\sum_{i=1}^n m(\boldsymbol{X}_i;\boldsymbol{\xi})$ *for* $m(\boldsymbol{X};\boldsymbol{\xi})=\boldsymbol{\xi}^T\boldsymbol{k}(\boldsymbol{X})$ *an estimator of* μ_0*, with* $\boldsymbol{k}(\boldsymbol{X})=(1,X_1,\ldots,X_p)^T$*. For fixed* $\boldsymbol{\xi}$*, the influence function of* $\hat{\mu}_{n,IMP}(\boldsymbol{\xi})$ *is* $\phi_{IMP}(\boldsymbol{O};\mu_0,\boldsymbol{\xi})=m(\boldsymbol{X};\boldsymbol{\xi})-\mu_0$ *and the squared bias is* $E^2\{m(\boldsymbol{X};\boldsymbol{\xi})-\mu_0\}$*. In this case,* $E\{\partial\phi_{IMP}(\boldsymbol{O};\mu_0,\boldsymbol{\xi})/\partial\boldsymbol{\xi}\}=E\{\boldsymbol{k}(\boldsymbol{X})\}$ *and this does not depend on* $\boldsymbol{\xi}$*. Clearly,* $\partial\phi_{IMP}(\boldsymbol{O};\mu_0,\boldsymbol{\xi})/\partial\boldsymbol{\xi}$ *does not provide an unbiased estimating function.*

4.2.2 Further properties

It follows from Proposition 3.1 (and Proposition 3.2 in Appendix 3.A) that at $\mathcal{M}(\boldsymbol{\psi})\cap\mathcal{M}(\boldsymbol{\xi})$, small perturbations (of the order one over root-n) in $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ do not affect the first-order asymptotic behavior of the doubly robust estimator in the sense that $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi})/\partial\boldsymbol{\xi}\}=\boldsymbol{0}$ for all $\boldsymbol{\xi}$ and $E\{\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi}^*)/\partial\boldsymbol{\psi}\}=\boldsymbol{0}$ for all $\boldsymbol{\psi}$. This local robustness is lost as soon as one of the working models is misspecified. The estimators $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{BR}}$ for the nuisance parameters are designed

to restore this local robustness property under model misspecification. It is hence not surprising that (the probability limits of) these estimators locally minimize $\text{bias}^2(\boldsymbol{\psi}, \boldsymbol{\xi}; \mu_0)$. Specifically, they ensure that the bias-reduced doubly robust estimator $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}})$ is first-order ancillary (Cox 1980) under misspecification of the working models in the sense formalized in the following corollary.

Corollary 4.1. *Under suitable regularity conditions (Robins et al. 1994, app. B),*

$$n^{1/2}\{\hat{\mu}_{n,DR}(\hat{\boldsymbol{\psi}}_n^{BR}, \hat{\boldsymbol{\xi}}_n^{BR}) - \mu_0\} = n^{-1/2}\sum_{i=1}^{n}\phi(\boldsymbol{O}_i; \mu_0, \boldsymbol{\psi}_{BR}^*, \boldsymbol{\xi}_{BR}^*) + o_p(1)$$

when $\hat{\boldsymbol{\psi}}_n^{BR}$ and $\hat{\boldsymbol{\xi}}_n^{BR}$ are the solutions to (4.3) and (4.4) with $\boldsymbol{\psi}_{BR}^* = plim(\hat{\boldsymbol{\psi}}_n^{BR})$ and $\boldsymbol{\xi}_{BR}^* = plim(\hat{\boldsymbol{\xi}}_n^{BR})$.

Proof. This follows from the proof of Proposition 3.1 because (under standard regularity conditions) $(\hat{\boldsymbol{\psi}}_n^{\text{BR}} - \boldsymbol{\psi}_{\text{BR}}^*)$ and $(\hat{\boldsymbol{\xi}}_n^{\text{BR}} - \boldsymbol{\xi}_{\text{BR}}^*)$ are $O_p(n^{-1/2})$ (see also Theorem 3.13 in Robins et al. (2008)). □

This first-order ancillarity implies that the first-order asymptotic behavior of $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}})$ is the same as that of $\hat{\mu}_{n,\text{DR}}(\boldsymbol{\psi}_{\text{BR}}^*, \boldsymbol{\xi}_{\text{BR}}^*)$, in which $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{BR}}$ are substituted by their probability limits $\boldsymbol{\psi}_{\text{BR}}^*$ and $\boldsymbol{\xi}_{\text{BR}}^*$, respectively. This has a number of important consequences. First, the asymptotic variance of the doubly robust estimator $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}})$ can be straightforwardly estimated as one over n times the sample variance of the values $\phi\{\boldsymbol{O}_i; \hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}}), \hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}}\}$, without having to correct for the estimation of the nuisance parameters. Similarly, a score test of the null hypothesis that $\mu = \tilde{\mu}$ for some $\tilde{\mu}$ simplifies to a one-sample t-test of the null hypothesis that $\phi(\boldsymbol{O}; \tilde{\mu}, \boldsymbol{\psi}_0, \boldsymbol{\xi}_0)$ has mean zero. Second, the estimators $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{BR}}$ tend to deliver reasonably efficient doubly robust estimators as will be confirmed in simulation studies in Section 4.4. This can be intuitively expected because an estimator tends to be less variable when evaluated at fixed nuisance parameter values instead of estimated ones. However, an efficiency benefit relative to the use of maximum likelihood estimation of the nuisance parameters is not theoretically guaranteed. One reason for this is that under model misspecifi-

cation, different estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\xi}}_n$ of the nuisance parameters may converge to different probability limits $\boldsymbol{\psi}^*$ and $\boldsymbol{\xi}^*$ and thereby influence the variance of $\phi(\boldsymbol{O}; \mu^*, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*)$, with μ^* the corresponding probability limit of the doubly robust estimator under model misspecification. A second reason is that an estimator may sometimes vary less when evaluated at estimated rather than known nuisance parameters (Pierce 1982; Rotnitzky et al. 2010).

4

4.2.3 Connection to the theory of *higher-order influence functions*

James Robins, Andrea Rotnitzky and Eric Tchetgen Tchetgen noted that Robins et al. (2008) use similar estimating equations like (4.3) and (4.4) in an intermediate simplifying step in the construction of higher-order influence functions, but with a different objective. In their approach, these estimating equations are not used to directly estimate nuisance parameters describing parametric working models. Instead they first obtain (potentially highly data-adaptive) initial estimators of these working models using a split sample. In this manner, these initial estimators are independent of the estimator for the target parameter and hence one does not need to adjust for its estimation. Robins et al. (2008) then use estimating equations similar to (4.3) and (4.4) to estimate nuisance parameters describing specific linear extensions of these initial estimators (where the dimension can increase with the sample size), estimated via the sample that is also used to estimate the target parameter; in contrast, we allow for arbitrary but finite-dimensional nuisance working models. As a result, first-order ancillarity (see our Corollary 4.1 and Theorem 3.13 in Robins et al. (2008)) with respect to their fluctuation parameters is obtained, which simplifies the derivation of higher-order influence functions. Our Theorem 4.2 is also similar to their Lemma 3 but theirs allows for infinite-dimensional nuisance parameters. No other advantage of doing so is mentioned in that paper.

4.3 Illustration: Missing Data Problem

4.3.1 Bias-reduced doubly robust estimator

To illustrate the bias-reduced doubly robust estimation strategy, we return to the missing data problem introduced in Section 3.3. From the influence function (3.5) of the doubly robust estimator, it follows that (4.3) equals

$$\sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})} \right\} m_{\boldsymbol{\xi}}(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})] = \mathbf{0}.$$

For instance, for $m(\boldsymbol{X}; \boldsymbol{\xi}) = \boldsymbol{\xi}^T \boldsymbol{k}(\boldsymbol{X})$ with $\boldsymbol{k}(\boldsymbol{X}) = (1, X_1, \ldots, X_p)^T$, this becomes

$$\sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})} \right\} \boldsymbol{k}(\boldsymbol{X}_i) = \mathbf{0}. \tag{4.5}$$

Here, the first equation (that is, by taking the first component of $\boldsymbol{k}(\boldsymbol{X})$), $n = \sum_{i=1}^{n} \{R_i / \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\}$, ensures that the inverse weights sum to the sample size; $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ then equals

$$n^{-1} \sum_{i=1}^{n} m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}) + \frac{n^{-1} \sum_{i=1}^{n} R_i / \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}) \{Y_i - m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\}}{n^{-1} \sum_{i=1}^{n} R_i / \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})},$$

also considered in Robins et al. (2007). The remaining equations in (4.5) impose that the sample mean of the covariates, $n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$, equals the weighted sample mean $n^{-1} \sum_{i=1}^{n} R_i \boldsymbol{X}_i / \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})$. These restrictions help to ensure stable weights; the bias-reduced doubly robust estimation strategy might therefore alleviate the problem of inefficiency due to highly variable weights (Robins et al. 2007). Restrictions (4.5) are also known as the calibration equations in the survey sampling literature (Särndal et al. 1989; Deville and Särndal 1992) where they are used to improve the simple Horvitz-Thompson estimator by making it unbiased under a linear prediction model (Kott and Liao 2012). See also Lumley et al. (2011) for a review of connections between doubly robust and calibration estimators. For linear outcome models (or more generally whenever $m(\boldsymbol{X}; \boldsymbol{\xi})$ lies within the span of the gradient $m_{\boldsymbol{\xi}}(\boldsymbol{X}, \boldsymbol{\xi})$), it

now follows from (4.5) that

$$\begin{aligned}\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}) &= n^{-1}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})} \\ &\qquad + \hat{\boldsymbol{\xi}}_n^{\mathrm{BR},T}\left[n^{-1}\sum_{i=1}^{n}\left\{1-\frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}\right\}\boldsymbol{k}(\boldsymbol{X}_i)\right] \\ &= n^{-1}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})} = \frac{n^{-1}\sum_{i=1}^{n}R_iY_i/\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}{n^{-1}\sum_{i=1}^{n}R_i/\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}.\end{aligned}$$

4

This demonstrates that the bias-reduced doubly robust estimator remarkably reduces to a simple IPTW estimator, making $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ sample bounded (Robins et al. 2007, sec. 4.1) in the sense that it lies with probability one within the admissable range of observed outcome values (i.e., $[Y_{\min},Y_{\max}]$, where $Y_{\min}=\min\{Y_i:R_i=1\}$ and $Y_{\max}=\max\{Y_i:R_i=1\}$) whenever the outcome Y is continuous with conditional mean linear in $\boldsymbol{X}$ (or $m(\boldsymbol{X};\boldsymbol{\xi})$ lies within the span of the gradient $m_{\boldsymbol{\xi}}(\boldsymbol{X},\boldsymbol{\xi})$). This is because the estimator

$$n^{-1}\sum_{i=1}^{n}\frac{R_i/\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}{n^{-1}\sum_{i=1}^{n}R_i/\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}Y_i$$

is a convex combination of the observed outcomes.

From (3.5), it now follows that the estimating equation (4.4) for $\boldsymbol{\xi}$ equals

$$\sum_{i=1}^{n}\left[\frac{R_i}{\pi^2(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}\{Y_i-m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\}\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\right]=\boldsymbol{0}.$$

For instance, when $\pi(\boldsymbol{X};\boldsymbol{\psi})$ is the logistic regression model expit$\{\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X})\}$ with $\boldsymbol{l}(\boldsymbol{X})=(1,X_1,\ldots,X_p)$, this becomes

$$\sum_{i=1}^{n}\left[R_i\frac{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}\left\{Y_i-m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\right\}\boldsymbol{l}(\boldsymbol{X}_i)\right]=\boldsymbol{0}, \tag{4.6}$$

which amounts to weighted least squares based on the complete cases with weights $\{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\}/\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})$. High (low) weights are thus given to covariate regions with low (high) probability of observing the outcome, thereby forcing the

model to fit well in regions with most missing data. Because by the first component of the estimating equation (4.6),

$$\sum_{i=1}^{n} \frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}\left\{Y_i - m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\right\} = \sum_{i=1}^{n} R_i\left\{Y_i - m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\right\},$$

the bias-reduced doubly robust estimator can now be equivalently written as a regression imputation estimator

$$\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}}) = n^{-1}\sum_{i=1}^{n}\{R_iY_i + (1-R_i)m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\},$$

which averages the observed outcome for responders and a predicted outcome for non-responders. This is desirable as it ensures that the aforementioned boundedness property is also effective whenever the outcome predictions obey the admissable range of the data. For instance, when Y is binary and $m(\boldsymbol{X};\boldsymbol{\xi}) = \mathrm{expit}\{\boldsymbol{\xi}^T\boldsymbol{k}(\boldsymbol{X})\}$, $m(\boldsymbol{X};\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ falls between zero and one so that $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ is guaranteed to lie between zero and one.

While the estimating equation (4.6) for $\boldsymbol{\xi}$ can be easily solved using weighted regression in standard statistical software, this is not so for the estimating equation (4.5) for $\boldsymbol{\psi}$. To accommodate this, arguing along the lines of Tan (2010), we define the function

$$\mathcal{F}(\boldsymbol{\psi}) = n^{-1}\sum_{i=1}^{n}\left[-R_i\exp\{-\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i)\} - (1-R_i)\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i)\right], \tag{4.7}$$

which is an integrated form of (4.5) in the sense that $\partial\mathcal{F}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ equals (4.5). The function $\mathcal{F}(\boldsymbol{\psi})$ is always concave and bounded on every bounded set for $\boldsymbol{\psi}$. The proposition below, which is an adaptation of the proof of Condition (12) in Tan (2010), gives a condition under which $\mathcal{F}(\boldsymbol{\psi})$ has a unique maximum.

Proposition 4.1. *The function $\mathcal{F}(\boldsymbol{\psi})$ is strictly concave and bounded from above if and only if the set $\Gamma_2 = \{\boldsymbol{\psi} \neq \mathbf{0} : \boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i) \geq 0$ for all $i = 1,\ldots,n$ with $R_i = 1$, and $n^{-1}\sum_{i=1}^{n}\{(1-R_i)\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i)\} \leq 0\}$ is empty.*

Proof. The function $\mathcal{F}(\boldsymbol{\psi})$ is bounded on every bounded set of vectors $\boldsymbol{\psi}$. More-

over, $\mathcal{F}(\boldsymbol{\psi})$ is a concave function because its Hessian

$$\mathcal{H}(\boldsymbol{\psi}) = n^{-1}\sum_{i=1}^{n}\left[-R_i \exp\left\{-\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X}_i)\right\}\boldsymbol{l}(\boldsymbol{X}_i)\boldsymbol{l}^T(\boldsymbol{X}_i)\right]$$

is semi negative definite for every choice of $\boldsymbol{\psi}$: for every vector $\boldsymbol{\omega}$, we thus have $\boldsymbol{\omega}^T\mathcal{H}(\boldsymbol{\psi})\boldsymbol{\omega} = n^{-1}\sum_{i=1}^{n}\left[-R_i\exp\{-\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i)\}\{\boldsymbol{\omega}^T\boldsymbol{l}(\boldsymbol{X}_i)\}^2\right] \leq 0$. Furthermore, if $\mathcal{H}(\boldsymbol{\psi})$ is negative definite, then $\mathcal{F}(\boldsymbol{\psi})$ is strictly concave. This will be the case if the set $\Gamma_1 = \{\boldsymbol{\psi} \neq \mathbf{0} : \boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i) = 0 \text{ for all } i = 1,\ldots,n \text{ with } R_i = 1\}$ is empty. Otherwise, there exists an $\boldsymbol{\omega} \neq \mathbf{0}$ such that $\boldsymbol{\omega}^T\mathcal{H}(\boldsymbol{\psi})\boldsymbol{\omega} = 0$ implying that $\boldsymbol{\omega}^T\boldsymbol{l}(\boldsymbol{X}_i) = 0$ for all $i = 1,\ldots,n$ with $R_i = 1$, which would indicate Γ_1 is not empty, a contradiction. Also, if there is an $\boldsymbol{\omega} \in \Gamma_1$, then $\mathcal{F}(\boldsymbol{\psi} + c\boldsymbol{\omega}) = \mathcal{F}(\boldsymbol{\psi}) - cn^{-1}\sum_{i=1}^{n}\{(1-R_i)\boldsymbol{\omega}^T\boldsymbol{l}(\boldsymbol{X}_i)\}$ is linear in $c \in \mathbb{R}$ implying $\mathcal{F}$ cannot be strictly concave. Thus, $\mathcal{F}$ is strictly concave if and only if Γ_1 is empty.

Next define the set $\Gamma_2 = \{\boldsymbol{\psi} \neq \mathbf{0} : \boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i) \geq 0 \text{ for all } i = 1,\ldots,n \text{ with } R_i = 1, \text{ and } n^{-1}\sum_{i=1}^{n}\{(1-R_i)\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X}_i)\} \leq 0\}$. The function $\mathcal{F}(\boldsymbol{\psi})$ will be bounded from above if and only if the set Γ_2 is empty. First we show that if Γ_2 is empty, than $\mathcal{F}(\boldsymbol{\psi})$ is bounded from above. We already know that this function is bounded on every finite subset. Suppose $\mathcal{F}$ is not bounded. Then there exists a sequence $(c_k, \boldsymbol{\omega}_k)$ where $c_k > 0$ and $\boldsymbol{\omega}_k$ is a unit vector such that $\mathcal{F}(c_k\boldsymbol{\omega}_k) \to \infty$ as $k \to \infty$ and thus $c_k \to \infty$. By compactness of the unit ball, there exists a subsequence (for which we use the same notation) such that $\boldsymbol{\omega}_k \to \boldsymbol{\omega}_0$ as $k \to \infty$. Then we must have that $\boldsymbol{\omega}_0^T\boldsymbol{l}(\boldsymbol{X}_i) \geq 0$ for all $i = 1,\ldots,n$ with $R_i = 1$, since otherwise, for some i with $R_i = 1$ we have that $\boldsymbol{\omega}_0^T\boldsymbol{l}(\boldsymbol{X}_i) < 0$ and hence that for k sufficiently large, $\boldsymbol{\omega}_k^T\boldsymbol{l}(\boldsymbol{X}_i) < 0$. Consequently, $\mathcal{F}(c_k\boldsymbol{\omega}_k) \to -\infty$, a contradiction. Additionally, $n^{-1}\sum_{i=1}^{n}\{(1-R_i)\boldsymbol{\omega}_0^T\boldsymbol{l}(\boldsymbol{X}_i)\} \geq 0$. Otherwise, if $n^{-1}\sum_{i=1}^{n}\{(1-R_i)\boldsymbol{\omega}_0^T\boldsymbol{l}(\boldsymbol{X}_i)\} > 0$, for k sufficiently large we have $n^{-1}\sum_{i=1}^{n}\{(1-R_i)\boldsymbol{\omega}_k^T\boldsymbol{l}(\boldsymbol{X}_i)\} > 0$, and hence $\mathcal{F}(c_k\boldsymbol{\omega}_k) \to -\infty$, a contradiction. Finally we show that if $\mathcal{F}(\boldsymbol{\psi})$ is bounded from above, then Γ_2 must be empty. If there exists an $\boldsymbol{\psi} \in \Gamma_2$, then $\mathcal{F}(\boldsymbol{\psi} + c\boldsymbol{\omega}) \to \infty$ if $c \to \infty$. □

From this proposition, it follows that if Γ_2 is empty, the function $\mathcal{F}(\boldsymbol{\psi})$ has a unique maximum which equals $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$. In Appendix 4.C, we provide an R-function to obtain the bias-reduced doubly robust estimator $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}})$ for a logistic propensity score model and both a linear and logistic regression model for the

conditional outcome.

The bias reduction promised by Theorem 4.1 may be substantial, as we argue next. Let $\delta(\boldsymbol{X};\boldsymbol{\psi}^*) = \pi(\boldsymbol{X};\boldsymbol{\psi}^*) - \pi_0(\boldsymbol{X})$ denote the degree of model misspecification in the model for the propensity score at given $\boldsymbol{X}$ and $\Delta(\boldsymbol{X};\boldsymbol{\xi}^*) = m(\boldsymbol{X};\boldsymbol{\xi}^*) - m_0(\boldsymbol{X})$ the degree of model misspecification in the working model for the conditional mean outcome at given $\boldsymbol{X}$. When both working models are misspecified, the asymptotic bias of the doubly robust estimator can be written as (see e.g., Vansteelandt et al. 2012)

$$\text{bias}(\boldsymbol{\psi}^*, \boldsymbol{\xi}^*; \mu_0) = E\left[\frac{\delta(\boldsymbol{X};\boldsymbol{\psi}^*)\Delta(\boldsymbol{X};\boldsymbol{\xi}^*)}{\pi(\boldsymbol{X};\boldsymbol{\psi}^*)}\right]. \tag{4.8}$$

It is thus driven by the degree of misspecification $\delta(\boldsymbol{X};\boldsymbol{\psi}^*)$ and $\Delta(\boldsymbol{X};\boldsymbol{\xi}^*)$ but may get inflated in regions with small propensity score. This inflation is a legitimate concern because in these regions with low propensity score, the probability of observing Y is low and misspecification in $m(\boldsymbol{X};\boldsymbol{\xi})$ is most likely. Bias-reduced doubly robust estimation prevents such inflation. For instance, using the first component of the vector of estimating equations in (4.6), we obtain that $E\left[\pi_0(\boldsymbol{X})\Delta(\boldsymbol{X};\boldsymbol{\xi}^*_{\text{BR}})/\pi(\boldsymbol{X};\boldsymbol{\psi}^*_{\text{BR}})\right] = E\{\Delta(\boldsymbol{X};\boldsymbol{\xi}^*_{\text{BR}})\pi_0(\boldsymbol{X})\}$. This is so whenever the logistic regression model for the propensity score includes an intercept. The asymptotic bias (4.8) can then be equivalently written as

$$E[\Delta(\boldsymbol{X};\boldsymbol{\xi}^*_{\text{BR}})\{1-\pi_0(\boldsymbol{X})\}],$$

and hence does not get severely inflated in covariate regions with small propensity score.

4.3.2 Graphical illustration

To gain some insight in the bias-reduced doubly robust estimation principle as compared to standard MLE, we visualize the bias reduction property using two simple examples. For both examples, consider a sample $\mathcal{O}$ of size $n = 10^5$, which we take to be large so that we can ignore sampling variability.

Example 1

For each individual $i = 1, \ldots, n$, we let

$$\begin{aligned} X_i &\stackrel{d}{=} N(0,1), \\ R_i|X_i &\stackrel{d}{=} \text{Ber}\{\text{expit}(-1+X_i^3)\} \text{ and} \\ Y_i|X_i &\stackrel{d}{=} N(X_i^2,1). \end{aligned}$$

The true mean outcome equals $\mu_0 = 1$. Misspecified (one-dimensional) working models are of the form $\pi(X;\psi) = \text{expit}(\psi X)$ and $m(X;\xi) = \xi X$. We obtain $\hat{\psi}_n^{\text{MLE}} = 1.115$ and $\hat{\xi}_n^{\text{MLE}} = 1.606$, leading to $\hat{\mu}_{n,\text{MLE}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE}}) = 0.237$ and consequently $\text{bias}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE}}; 1) = -0.763$. The bias-reduced doubly robust estimation strategy yields $\hat{\psi}_n^{\text{BR}} = 2.254$ and $\hat{\xi}_n^{\text{BR}} = -0.959$, leading to $\hat{\psi}_{n,\text{BR}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\psi}_n^{\text{BR}}, \hat{\xi}_n^{\text{BR}}) = 0.663$ and a smaller bias of $\text{bias}(\hat{\psi}_n^{\text{BR}}, \hat{\xi}_n^{\text{BR}}; 1) = -0.337$, as theoretically expected. Figure 4.1 shows a contour plot of the log of the squared first-order asymptotic bias as a function of the nuisance parameters ψ and ξ. Darker values indicate smaller bias. The cross '$\times$' indicates the point $(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE}}) = (1.115, 1.606)$, which approximates $(\psi^*_{\text{MLE}}, \xi^*_{\text{MLE}})$ and the bullet '$\bullet$' indicates the point $(\hat{\psi}_n^{\text{BR}}, \hat{\xi}_n^{\text{BR}}) = (2.254, -0.959)$, which approximates $(\psi^*_{\text{BR}}, \xi^*_{\text{BR}})$. Figure 4.1 illustrates the defining property of the bias-reduced estimation principle: the squared first-order asymptotic bias of the doubly robust estimator is locally minimized in the point $(2.254, -0.959) \approx (\psi^*_{\text{BR}}, \xi^*_{\text{BR}})$. Bias-reduction is indeed local, as we observe that there are two other regions where even smaller bias near zero is attained. The location of these regions depends strongly on the underlying data-generating mechanism, so that they cannot be defined by solving an estimating equation without further knowledge of the true data-generating mechanism. To illustrate that the location of such regions depends on the underlying data-generating mechanism, we show a similar plot for another example.

Example 2

In this second example, for each individual $i = 1, \ldots, n$, we let

$$X_i \stackrel{d}{=} N(1,1),$$

Figure 4.1: *Contourplot of the log of the squared first-order asymptotic bias* $\log\{bias^2(\psi,\xi;1)\}$ *as a function of the nuisance parameters* ψ *and* ξ *for* ***Example 1*** *with* $\times = (1.115, 1.606) \approx (\psi^*_{MLE}, \xi^*_{MLE})$ *and* $\bullet = (2.254, -0.959) \approx (\psi^*_{BR}, \xi^*_{BR})$.

$$R_i|X_i \stackrel{d}{=} \text{Ber}\{\text{expit}(-1+X_i^2)\} \text{ and}$$
$$Y_i|X_i \stackrel{d}{=} N(X_i^2, 1).$$

The true mean outcome equals $\mu_0 = 2$. Misspecified (one-dimensional) working models are also of the form $\pi(X;\psi) = \text{expit}(\psi X)$ and $m(X;\xi) = \xi X$. Figure 4.2 shows a contour plot of the log of the squared first-order asymptotic bias as a function of the nuisance parameters ψ and ξ, where darker values again indicate smaller bias. The cross '$\times$' now indicates the point $(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE}}) = (0.737, 2.157)$, which approximates $(\psi^*_{\text{MLE}}, \xi^*_{\text{MLE}})$, leading to $\hat{\mu}_{n,\text{MLE}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE}}) = 2.393$ and $\text{bias}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE}}; 2) = 0.393$. The bullet '$\bullet$' indicates the point $(\hat{\psi}_n^{\text{BR}}, \hat{\xi}_n^{\text{BR}}) = (0.609, 1.220)$, which approximates $(\psi^*_{\text{BR}}, \xi^*_{\text{BR}})$, leading to the bias-reduced doubly

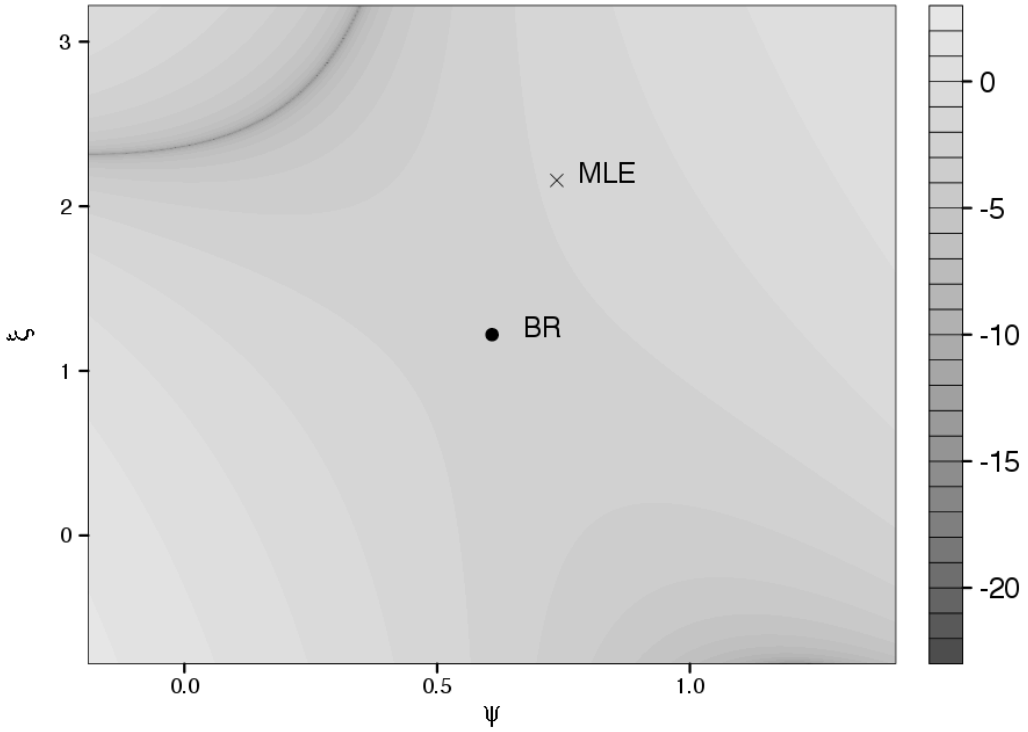


Figure 4.2: *Contourplot of the log of the squared first-order asymptotic bias* $\log\{bias^2(\psi,\xi;2)\}$ *as a function of the nuisance parameters* ψ *and* ξ *for* ***Example 2*** *with* $\times = (0.737, 2.157) \approx (\psi^*_{MLE}, \xi^*_{MLE})$ *and* $\bullet = (0.609, 1.220) \approx (\psi^*_{BR}, \xi^*_{BR})$.

robust estimator $\hat{\mu}_{n,\text{BR}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\psi}^{\text{BR}}_n, \hat{\xi}^{\text{BR}}_n) = 2.316$ and $\text{bias}(\hat{\psi}^{\text{BR}}_n, \hat{\xi}^{\text{BR}}_n; 2) = 0.316$, which is smaller than the bias of $\hat{\mu}_{n,\text{MLE}}$. The result in this second example (see Figure 4.2) is similar to the result in the first example (see Figure 4.1). However, do note that the regions where smaller bias near zero is attained are located in a different position, not trackable by solving an estimating equation without further knowledge of the true data-generating mechanism.

We have previously seen that small perturbations (of the order one over root-n) in the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ do not affect the first-order asymptotic behavior of the doubly robust estimator when the intersection model $\mathcal{M}(\boldsymbol{\psi}) \cap \mathcal{M}(\boldsymbol{\xi})$ holds. In Section 4.2.2, we found that besides reducing bias, the bias-reduced estimation procedure also extends the aforementioned property to working model misspecification. This is visually suggested by both Figure 4.1 and Figure 4.2 in

the sense that the estimators $\hat{\psi}_n^{\mathrm{BR}}$ and $\hat{\xi}_n^{\mathrm{BR}}$ are situated in a region where small perturbations of the nuisance parameters do not heavily affect the bias of the doubly robust estimator (in contrast to the regions that deliver near-zero bias, which are very unstable). This is shown formally in Corollary 4.1, which shows that the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\psi}_n^{\mathrm{BR}}, \hat{\xi}_n^{\mathrm{BR}})$ is first-order ancillary (Cox 1980) under misspecification of the working models.

4.3.3 Connection to the theory of *targeted estimation of nuisance parameters*

Like the bias-reduced doubly robust estimation procedure, a recent proposal by van der Laan (2014) also focuses on bias reduction. This proposal is different in that it is based on removing an approximation to the first-order bias of the doubly robust estimator by cleverly fitting highly data-adaptive working models, on which we elaborate below. This section can be skipped by the less interested reader.

Given a working model $\pi(\boldsymbol{X})$ for the true propensity score $\pi_0(\boldsymbol{X}) = P(R=1|\boldsymbol{X})$ and a working model $m(\boldsymbol{X})$ for the true conditional mean outcome $m_0(\boldsymbol{X}) = E(Y|\boldsymbol{X})$ (where both working models can be parametric models), the bias of the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\pi, m) = n^{-1}\sum_{i=1}^n R_i Y_i/\pi(\boldsymbol{X}_i) + \{1 - R_i/\pi(\boldsymbol{X}_i)\}\, m(\boldsymbol{X}_i)$ is given by

$$\mathrm{bias}(\pi, m; \mu_0) = E\left[\left\{\frac{\pi_0(\boldsymbol{X})}{\pi(\boldsymbol{X})} - 1\right\}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X})\}\right].$$

When we insert the probability limits $\pi^*(\boldsymbol{X})$ and $m^*(\boldsymbol{X})$ of estimators $\hat{\pi}_n(\boldsymbol{X})$ and $\hat{m}_n(\boldsymbol{X})$ for these working models that at least converge at an $n^{1/4+\delta}$-rate, with $\delta > 0$, this is the first-order bias (see also equation (4.8) for parametric working models). Both van der Laan (2014) and the bias-reduced doubly robust estimation procedure focus on this term, however from a different perspective.

For comparability and the sake of simplicity, for the moment, we restrict ourselves to parametric working models $\pi(\boldsymbol{X}; \boldsymbol{\psi})$ and $m(\boldsymbol{X}; \boldsymbol{\xi})$ for finite-dimensional parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$. If we would know $m_0(\boldsymbol{X})$, we could find parameter values for $\boldsymbol{\psi}$ such that

$$E\left[\left\{\frac{R}{\pi(\boldsymbol{X}; \boldsymbol{\psi})} - 1\right\}\{m_0(\boldsymbol{X}) - m(\boldsymbol{X}; \boldsymbol{\xi})\}\right] = 0.$$

In the context of the bias-reduced doubly robust estimation strategy, this could be accomplished by adding $m_0(\boldsymbol{X}) - m(\boldsymbol{X};\boldsymbol{\xi})$ as a covariate to the outcome model using for instance the identity-link. Similarly, if $\pi_0(\boldsymbol{X})$ would be known, we could find parameter values for $\boldsymbol{\xi}$ such that

$$E\left[\frac{R}{\pi_0(\boldsymbol{X})}\left\{\frac{\pi_0(\boldsymbol{X})}{\pi(\boldsymbol{X};\boldsymbol{\psi})}-1\right\}\{Y-m(\boldsymbol{X};\boldsymbol{\xi})\}\right]=0.$$

In the context of the bias-reduced doubly robust estimation strategy, this could be accomplished by adding $\frac{R}{\pi_0(\boldsymbol{X})}\left(\frac{\pi_0(\boldsymbol{X})-\pi(\boldsymbol{X};\boldsymbol{\psi})}{1-\pi(\boldsymbol{X};\boldsymbol{\psi})}\right)$ as a covariate to the propensity score model using the logit-link. Thus, if either $\pi_0(\boldsymbol{X})$ or $m_0(\boldsymbol{X})$ would be known, we could estimate the nuisance parameters such that this bias term is zero. Of course, these strategies are infeasible because neither $\pi_0(\boldsymbol{X})$ and $m_0(\boldsymbol{X})$ are known.

The bias-reduced doubly robust estimation strategy now aims to identify limiting values $\boldsymbol{\psi}^*_{\text{BR}}$ and $\boldsymbol{\xi}^*_{\text{BR}}$ of the nuisance parameters that locally minimize the square of this first-order bias term in the direction of these nuisance parameters under possible misspecification of **both** working models. In Section 4.2.1, we showed this can be accomplished by defining these limits solving (4.1) and (4.2). The corresponding nuisance parameter estimators $\hat{\boldsymbol{\psi}}^{\text{BR}}_n$ and $\hat{\boldsymbol{\xi}}^{\text{BR}}_n$ are defined as solutions to the empirical versions of these equations. Under correct working model specification, the probability limits $\boldsymbol{\psi}^*_{\text{BR}}$ and $\boldsymbol{\xi}^*_{\text{BR}}$ equal the truth. The relevant conditions we impose are thus

$$E\left[\left\{\frac{\pi_0(\boldsymbol{X})}{\pi(\boldsymbol{X};\boldsymbol{\psi}^*_{\text{BR}})}-1\right\}\frac{\partial m(\boldsymbol{X};\boldsymbol{\xi}^*_{\text{BR}})}{\partial\boldsymbol{\xi}}\right]=\mathbf{0} \tag{4.9}$$

and

$$E\left[\{m_0(\boldsymbol{X})-m(\boldsymbol{X};\boldsymbol{\xi})\}\frac{R}{\pi^2(\boldsymbol{X};\boldsymbol{\psi}^*_{\text{BR}})}\frac{\partial\pi(\boldsymbol{X};\boldsymbol{\psi}^*_{\text{BR}})}{\partial\boldsymbol{\psi}}\right]=\mathbf{0}. \tag{4.10}$$

In contrast, the proposal of van der Laan (2014) is to approximate this bias using reduced working models and then to adapt the previous infeasible strategy by targeting initial high-dimensional fits of the working models. Specifically, note that the bias$(\pi^*,m^*;\mu_0)$ can be written as $E\left[\{R-\pi^*(\boldsymbol{X})\}\{m_0(\boldsymbol{X})-m^*(\boldsymbol{X})\}/\pi^*(\boldsymbol{X})\right]$, so that this bias can be approximated by approximating the true $m_0(\boldsymbol{X})-\hat{m}_n(\boldsymbol{X})$ via the reduction $m^r_0(\boldsymbol{X})=m^r_0\{\hat{\pi}_n(\boldsymbol{X}),\hat{m}_n(\boldsymbol{X})\}=E\{Y-\hat{m}_n(\boldsymbol{X})|R=1,\hat{\pi}_n(\boldsymbol{X})\}$, which

is estimated by $\hat{m}_n^r(\boldsymbol{X})$ via nonparametric regression and is assumed to be consistent for $m_0^r(\boldsymbol{X})$. This approximation divided by the initial fit of the propensity score is then used to fluctuate the initial fit of the propensity score. To be specific, van der Laan (2014) proposes the fluctuation model

$$\operatorname{logit}\hat{\pi}_n(\varepsilon_R)(\boldsymbol{X}) = \operatorname{logit}\hat{\pi}_n(\boldsymbol{X}) + \varepsilon_R \frac{\hat{m}_n^r(\boldsymbol{X})}{\hat{\pi}_n(\boldsymbol{X})},$$

where the optimal fluctuation is defined by $\hat{\varepsilon}_{n,R} = \arg\min_{\varepsilon_R} \sum_{i=1}^n \mathcal{L}\{\hat{\pi}_n(\varepsilon_R)\}(\boldsymbol{O}_i)$, with $\mathcal{L}\{\pi(\varepsilon_R)\}(\boldsymbol{O}) = -[R\log\{\pi(\varepsilon_R)(\boldsymbol{X})\} + (1-R)\log\{1-\pi(\varepsilon_R)(\boldsymbol{X})\}]$. It follows that $\hat{\varepsilon}_{n,R}$ solves the score equation

$$\sum_{i=1}^n \{R_i - \hat{\pi}_n(\hat{\varepsilon}_{n,R})(\boldsymbol{X}_i)\}\frac{\hat{m}_n^r(\boldsymbol{X}_i)}{\hat{\pi}_n(\boldsymbol{X}_i)} = 0,$$

and the procedure is iterated until convergence. In the limit, an approximation of the bias term is put to zero:

$$E\left[\{R-\pi^*(\boldsymbol{X})\}\frac{m^{r,*}(\boldsymbol{X})}{\pi^*(\boldsymbol{X})}\right] = E\left[\left\{\frac{\pi_0(\boldsymbol{X})}{\pi^*(\boldsymbol{X})} - 1\right\} m^{r,*}(\boldsymbol{X})\right\} = 0 \tag{4.11}$$

with the * indicating the probability limits. Similarly, because the bias can be written as $E([E\{R|\pi^*(\boldsymbol{X}), m^*(\boldsymbol{X})\}/\pi^*(\boldsymbol{X}) - 1]\{Y - m^*(\boldsymbol{X})\})$, this bias can be approximated by approximating the true $\pi_0(\boldsymbol{X})$ via $\pi_0^r(\boldsymbol{X}) = \pi_0^r\{\hat{\pi}_n(\boldsymbol{X}), \hat{m}_n(\boldsymbol{X})\} = P\{R=1|\hat{\pi}_n(\boldsymbol{X}), \hat{m}_n(\boldsymbol{X})\}$, which is estimated by $\hat{\pi}_n^r(\boldsymbol{X})$ via nonparametric regression and is assumed to be consistent for $\pi_0^r(\boldsymbol{X})$. This approximation is then used in the construction of the covariate $\frac{1}{\hat{\pi}_n^r(\boldsymbol{X})}\left\{\frac{\hat{\pi}_n^r(\boldsymbol{X})}{\hat{\pi}_n(\boldsymbol{X})} - 1\right\}$, used to fluctuate the initial fit of the outcome model. Specifically, van der Laan (2014) proposes the fluctuation model

$$\operatorname{logit}\hat{m}_n(\varepsilon_Y)(\boldsymbol{X}) = \operatorname{logit}\hat{m}_n(\boldsymbol{X}) + \varepsilon_Y \frac{1}{\hat{\pi}_n^r(\boldsymbol{X})}\left\{\frac{\hat{\pi}_n^r(\boldsymbol{X})}{\hat{\pi}_n(\boldsymbol{X})} - 1\right\},$$

where the optimal fluctuation is defined by $\hat{\varepsilon}_{n,Y} = \arg\min_{\varepsilon_Y} \sum_{i=1}^n \mathcal{L}\{\hat{m}_n(\varepsilon_Y)\}(\boldsymbol{O}_i)$, with $\mathcal{L}\{m(\varepsilon_Y)\}(\boldsymbol{O}) = -R[Y\log\{m(\varepsilon_Y)(\boldsymbol{X})\} + (1-Y)\log\{1-m(\varepsilon_Y)(\boldsymbol{X})\}]$. It fol-

lows that $\hat{\varepsilon}_{n,Y}$ solves the score equation

$$\sum_{i=1}^{n}\{Y_i - \hat{m}_n(\hat{\varepsilon}_{n,Y})(\boldsymbol{X}_i)\}\frac{R_i}{\hat{\pi}_n^r(\boldsymbol{X}_i)}\left\{\frac{\hat{\pi}_n^r(\boldsymbol{X}_i)}{\hat{\pi}_n(\boldsymbol{X}_i)} - 1\right\},$$

and the procedure is iterated until convergence. In the limit, an approximation of the bias term is put to zero:

$$\begin{aligned} &E\left[\frac{R}{\pi^{r,*}(\boldsymbol{X})}\left\{\frac{\pi^{r,*}(\boldsymbol{X})}{\pi^*(\boldsymbol{X})} - 1\right\}\{Y - m^*(\boldsymbol{X})\}\right] \\ &\quad = E\left[\frac{R}{\pi^{r,*}(\boldsymbol{X})}\left\{\frac{\pi^{r,*}(\boldsymbol{X})}{\pi^*(\boldsymbol{X})} - 1\right\}\{m_0(\boldsymbol{X}) - m^*(\boldsymbol{X})\}\right] = 0. \end{aligned} \tag{4.12}$$

We believe that how much bias-reduction of the first-order bias is accomplished heavily depends on the quality of the approximations $\pi^{r,*}(\boldsymbol{X})$ and $m^{r,*}(\boldsymbol{X})$. For a detailed presentation of these methods, we refer to van der Laan (2014).

Comparing (4.9) with (4.11) shows that both methodologies imply different restrictions and hence, one does not imply the other. Moreover, suppose that $\pi(\boldsymbol{X};\boldsymbol{\psi}^*_{\text{BR}})$ would equal the limit $\pi^*(\boldsymbol{X})$, then (4.9) even implies (4.11) when $m^{r,*}(\boldsymbol{X})$ happens to belong to the linear span of $\partial m(\boldsymbol{X};\boldsymbol{\xi}^*_{\text{BR}})/\partial\boldsymbol{\xi}$, e.g., (although unrealistic) both $m(\boldsymbol{X};\boldsymbol{\xi}^*_{\text{BR}})$ and $\pi(\boldsymbol{X};\boldsymbol{\psi}^*_{\text{BR}})$ are linear in the covariates and $m^{r,*}(\boldsymbol{X})$ is a simple linear regression on $\pi^*(\boldsymbol{X})$. However, vice versa, the restrictions imposed by (4.11) do not imply those imposed by (4.9).

In view of simplicity, the mainstream use of parametric models and the difficulty of obtaining good approximations to the bias, the bias-reduced doubly robust estimation procedure avoids such approximations and focuses on bias reduction under misspecification of (both) parametric working models.

4.3.4 Other alternatives

In Section 4.4, we will compare the performance of the bias-reduced doubly robust estimator for the mean outcome susceptible to missingness with several other alternatives proposed in the literature. Below, we briefly describe each of these alternatives.

Bounded and efficient doubly robust estimation with inverse weighting

Tan (2010) proposes a calibrated likelihood estimator, extending the results from Tan (2006). The proposed doubly robust estimators have desirable efficiency and boundedness properties in the sense it is a doubly robust sample bounded IPTW estimator with enhanced efficiency under a correctly specified propensity score model. The calibrated likelihood estimator can obtained via the `iWeigReg` package for R available on CRAN (Tan and Shu 2013), which is also used in the simulation studies in Section 4.4. Below, we briefly outline this procedure when the estimated propensity score is treated as known, as suggested in Tan and Shu (2013), which avoids unreliable estimates due to multicollinearity in the extended propensity score model (which is also discussed in Tan (2006)). For a detailed presentation of these methods, we refer to Tan (2010).

Consider the working models $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for $P(R=1|\boldsymbol{X})$ and $m(\boldsymbol{X};\boldsymbol{\xi})$ for $E(Y|X)$ and obtain estimators via MLE; $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{MLE}}$ (e.g., by solving estimating equations (2.30) and (2.32)). Define $\hat{\boldsymbol{\upsilon}}_n(\boldsymbol{X}) = \{1, m(\boldsymbol{X};\hat{\boldsymbol{\xi}}_n^{\text{MLE}})\}^T$ and let $\hat{\boldsymbol{h}}_n(\boldsymbol{X}) = \{1-\pi(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\text{MLE}})\}\hat{\boldsymbol{\upsilon}}_n(\boldsymbol{X})$. Next define the linear extended propensity score model

$$\omega(\boldsymbol{X};\boldsymbol{\lambda}) = \pi(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\text{MLE}}) + \boldsymbol{\lambda}^T\hat{\boldsymbol{h}}_n(\boldsymbol{X}). \tag{4.13}$$

The calibrated likelihood procedure now proceeds by calibrating the coefficients $\boldsymbol{\lambda}$ in the linear extended propensity score model (4.13). This is accomplished by two steps of maximizing concave functions.

1. Compute the estimator $\hat{\boldsymbol{\lambda}}_n^{(1)}$ which is defined as $\hat{\boldsymbol{\lambda}}_n^{(1)} = \arg\max_{\boldsymbol{\lambda}} \kappa_1(\boldsymbol{\lambda})$, with

$$\kappa_1(\boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^{n}\left[R_i\log\{\omega(\boldsymbol{X}_i;\boldsymbol{\lambda})\} + (1-R_i)\log\{1-\omega(\boldsymbol{X}_i;\boldsymbol{\lambda})\}\right].$$

The function $\kappa_1(\boldsymbol{\lambda})$ is finite and concave on the set

$$\left\{\boldsymbol{\lambda} : \omega(\boldsymbol{X}_i;\boldsymbol{\lambda}) > 0 \text{ if } R_i = 1 \text{ and } \omega(\boldsymbol{X}_i;\boldsymbol{\lambda}) < 1 \text{ if } R_i = 0,\, i = 1,\ldots,n\right\}$$

and is moreover strictly concave and bounded from above if and only the set

$$\left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^T \hat{\boldsymbol{h}}_n(\boldsymbol{X}) \geq 0 \text{ if } R_i = 1 \text{ and } \boldsymbol{\lambda}^T \hat{\boldsymbol{h}}_n(\boldsymbol{X}) \leq 0 \text{ if } R_i = 0,\, i = 1,\ldots,n\right\}$$

is empty, under which $\kappa_1(\boldsymbol{\lambda})$ has a unique maximum.

2. Compute the estimator $\hat{\boldsymbol{\lambda}}_n^{(2)}$ which is defined as $\hat{\boldsymbol{\lambda}}_n^{(2)} = \arg\max_{\boldsymbol{\lambda}} \kappa_2(\boldsymbol{\lambda})$, with

$$\kappa_2(\boldsymbol{\lambda}) = n^{-1} \sum_{i=1}^{n} \left[R_i \frac{\log\{\omega(\boldsymbol{X}_i;\boldsymbol{\lambda})\} - \log\{\omega(\boldsymbol{X}_i;\hat{\boldsymbol{\lambda}}_n^{(1)})\}}{1 - \pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE}})} - \boldsymbol{\lambda}_1^T \hat{\boldsymbol{\upsilon}}_n(\boldsymbol{X}_i) \right].$$

The function $\kappa_2(\boldsymbol{\lambda})$ is finite and concave on the set

$$\left\{\boldsymbol{\lambda} : \omega(\boldsymbol{X}_i;\boldsymbol{\lambda}) > 0 \text{ if } R_i = 1,\, i = 1,\ldots,n\right\}$$

and is moreover strictly concave and bounded from above if and only the set

$$\left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^T \hat{\boldsymbol{\upsilon}}_n(\boldsymbol{X}) \geq 0 \text{ if } R_i = 1,\, i = 1,\ldots,n \text{ and } n^{-1} \sum_{i=1}^{n} \boldsymbol{\lambda}^T \hat{\boldsymbol{\upsilon}}_n(\boldsymbol{X}_i) \leq 0\right\}$$

is empty, under which $\kappa_2(\boldsymbol{\lambda})$ has a unique maximum. This implies that

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \left[\left\{ \frac{R_i}{\omega(\boldsymbol{X}_i;\hat{\boldsymbol{\lambda}}_n^{(2)})} - 1 \right\} \hat{\boldsymbol{\upsilon}}_n(\boldsymbol{X}_i) \right], \tag{4.14}$$

implying that $1 = n^{-1}\sum_{i=1}^{n} R_i/\omega(\boldsymbol{X}_i;\hat{\boldsymbol{\lambda}}_n^{(2)})$ and that $n^{-1}\sum_{i=1}^{n} m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{MLE}}) = n^{-1}\sum_{i=1}^{n} R_i m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{MLE}})/\omega(\boldsymbol{X}_i;\hat{\boldsymbol{\lambda}}_n^{(2)})$. These constraints are similar to those in (4.5) of Section 4.3 but those implied in (4.14) are for the linear extended propensity score model while those in (4.5) are concerning the initial propensity score model. Note that when the initial propensity score model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ is correctly specified, $\hat{\boldsymbol{\lambda}}_n^{(2)} \xrightarrow{p} \mathbf{0}$.

From this, one obtains the doubly robust calibrated likelihood estimator $\hat{\mu}_{n,\mathrm{TAN}} = n^{-1}\sum_{i=1}^{n} R_i Y_i/\omega(\boldsymbol{X}_i;\hat{\boldsymbol{\lambda}}_n^{(2)})$, where the double robustness is guaranteed by (4.14).

Projection estimator

The construction of the projection estimation of Cao et al. (2009) proceeds under a correctly specified parametric propensity score model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for $P(R=1|\boldsymbol{X})$, where $\boldsymbol{\psi}$ is estimated via the MLE $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$, e.g., solving (2.30) for a logistic propensity score model. The parameter $\boldsymbol{\xi}$ indexing the model $m(\boldsymbol{X};\boldsymbol{\xi})$ for the conditional mean outcome $E(Y|\boldsymbol{X})$ is then estimated using the estimator $\hat{\boldsymbol{\xi}}_n^{\text{PROJ}}$ having the property that its probability limit $\boldsymbol{\xi}^*_{\text{PROJ}}$ minimizes the asymptotic variance of the doubly robust estimator using the parametric model $m(\boldsymbol{X};\boldsymbol{\xi})$, even under misspecification of $m(\boldsymbol{X};\boldsymbol{\xi})$ and that it equals $\boldsymbol{\xi}_0$ under correct specification of $m(\boldsymbol{X};\boldsymbol{\xi})$. This is accomplished by finding the value $\boldsymbol{\xi}^*_{\text{PROJ}}$ that minimizes the variance of

$$\frac{RY}{\pi_0(\boldsymbol{X})} - \frac{R-\pi_0(\boldsymbol{X})}{\pi_0(\boldsymbol{X})}\left\{m(\boldsymbol{X};\boldsymbol{\xi}) - \boldsymbol{c}^{*,T}(\boldsymbol{\xi})\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{1-\pi_0(\boldsymbol{X})}\right\} - \mu, \tag{4.15}$$

with

$$\begin{aligned}
\boldsymbol{c}^{*,T}(\boldsymbol{\xi}) &= -\Gamma_0^T(\boldsymbol{\xi})\Sigma^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi},0},\\
\Gamma_0(\boldsymbol{\xi}) &= E\left[\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\{m_0(\boldsymbol{X}) - m(\boldsymbol{X};\boldsymbol{\xi})\}/\pi_0(\boldsymbol{X})\right],\\
\Sigma_{\boldsymbol{\psi}\boldsymbol{\psi},0} &= E\left(\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\pi^T_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)/[\pi_0(\boldsymbol{X})\{1-\pi_0(\boldsymbol{X})\}]\right).
\end{aligned}$$

The variance of (4.15) equals

$$E\left[\frac{1-\pi_0(\boldsymbol{X})}{\pi_0(\boldsymbol{X})}\left\{Y - m(\boldsymbol{X};\boldsymbol{\xi}) - \boldsymbol{c}^{*,T}(\boldsymbol{\xi})\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{1-\pi_0(\boldsymbol{X})}\right\}\right] + \text{var}(Y). \tag{4.16}$$

The estimator $\hat{\boldsymbol{\xi}}_n^{\text{PROJ}}$ is then obtained by jointly solving $(\hat{\boldsymbol{\xi}}_n^{\text{PROJ},T}, \hat{\boldsymbol{c}}_n^{\text{PROJ},T})^T$ from the estimating equation

$$\begin{aligned}
\boldsymbol{0} = \sum_{i=1}^n R_i \frac{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}{\pi^2(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}
&\begin{bmatrix} m_{\boldsymbol{\xi}}\left(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\text{PROJ}}\right) \\ \dfrac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})} \end{bmatrix}\\
&\times\left\{Y_i - m\left(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\text{PROJ}}\right) - \hat{\boldsymbol{c}}_n^{\text{PROJ},T}\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}\right\}.
\end{aligned}$$

Under correct specification of the propensity score model, we know that $\hat{\boldsymbol{\psi}}_n^{\text{MLE}} \xrightarrow{p} \boldsymbol{\psi}_0$ and from Cao et al. (2009), if also follows that $\hat{\boldsymbol{\xi}}_n^{\text{PROJ}} \xrightarrow{p} \boldsymbol{\xi}_{\text{PROJ}}^*$ and $\hat{\boldsymbol{c}}_n^{\text{PROJ}} \xrightarrow{p} \boldsymbol{c}(\boldsymbol{\xi}_{\text{PROJ}}^*)$. Furthermore, when $m(\boldsymbol{X};\boldsymbol{\xi})$ is correctly specified but $\pi(\boldsymbol{X};\boldsymbol{\psi})$ is not, $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ will converge to some limit $\boldsymbol{\psi}^*$ but now $\hat{\boldsymbol{\xi}}_n^{\text{PROJ}} \xrightarrow{p} \boldsymbol{\xi}_0$ and $\hat{\boldsymbol{c}}_n^{\text{PROJ}} \xrightarrow{p} \boldsymbol{c}(\boldsymbol{\xi}_0) = \boldsymbol{0}$. From this, one obtains the projection estimator $\hat{\mu}_{n,\text{PROJ}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}, \hat{\boldsymbol{\xi}}_n^{\text{PROJ}})$. For more details, we refer to Cao et al. (2009).

4

Generalized Boosted Models

The alternative doubly robust estimator, introduced in Ridgeway and McCaffrey (2007), differs from the other alternatives considered in the simulation studies from Section 4.4 in that it does not rely on a parametric logistic regression model for the propensity score. Instead, it uses a generalized boosted model (GBM, see McCaffrey et al. (2004)), a multivariate nonparametric regression technique, to estimate the propensity score.

GBMs, first introduced in Ridgeway (1999), are automated and data-adaptive algorithms based on iteratively forming a collection of simple regression trees (Breiman et al. 1984) and adding them together to estimate the propensity score. Specifically, GBMs model the log-odds propensity score $g_0(\boldsymbol{X}) = \log[\pi_0(\boldsymbol{X})/\{1 - \pi_0(\boldsymbol{X})\}]$, which is unbounded, simplifying computations. The algorithm then proceeds as follows:

1. Initialize $\hat{g}_n^{(0)}(\boldsymbol{X}) = \log[n^{-1}\sum_{i=1}^n R_i/\{1 - n^{-1}\sum_{i=1}^n R_i\}]$.

2. For each m in $1,\ldots,M$ (with M the prespecified number of iterations), select a different random subsample (e.g., 50% of the observations). Then, do:

   (a) Let $\text{Res}_i = R_i - \text{expit}\{\hat{g}_n^{(m-1)}(\boldsymbol{X}_i)\}$, the residual of the current fit.

   (b) Construct a simple regression tree based on the residuals Res_i to partition the covariate space into terminal nodes $T_1,\ldots,T_K$.

   (c) For each of the terminal nodes T_k, $k = 1,\ldots,K$, compute the update

$$\theta_k = \frac{\sum_{\boldsymbol{X}_i \in T_k} \text{Res}_i}{\sum_{\boldsymbol{X}_i \in T_k} \text{expit}\{\hat{g}_n^{(m-1)}(\boldsymbol{X}_i)\}[1 - \text{expit}\{\hat{g}_n^{(m-1)}(\boldsymbol{X}_i)\}]}.$$

This update guarantees an improvement in the observed logistic log-likelihood.

(d) Update the logistic regression model as

$$\hat{g}_n^{(m)}(\boldsymbol{X}_i) \leftarrow \hat{g}_n^{(m-1)}(\boldsymbol{X}_i) + \alpha\theta_{k(\boldsymbol{X}_i)},$$

where $k(\boldsymbol{X}_i)$ indicates to which terminal node an observation with covariates $\boldsymbol{X}_i$ belongs and $\alpha \in (0,1]$ a shrinkage coefficient to reduce variability.

McCaffrey et al. (2004) suggest to allow trees with a maximum of four splits and to use $\alpha = 0.0005$ and $M = 20000$. GBMs for the propensity score can be fitted using the Toolkit for Weighting and Analysis of Nonequivalent Groups, the `twang` package for R available on CRAN (Ridgeway et al. 2015), which is used in the simulation studies in Section 4.4 using the tuning parameter values as suggested in McCaffrey et al. (2004), to which we also refer for more details concerning GBMs.

Targeted Maximum Likelihood Estimation (TMLE)

Targeted maximum likelihood estimation (and more generally, targeted minimum loss-based estimation), abbreviated TMLE, originally introduced in van der Laan and Rubin (2006), is a general method to obtain doubly robust substitution estimators of parameters in semi- and nonparametric causal inference and missing data models. Below, we briefly describe one version of TMLE for our particular problem of interest as presented in Gruber and van der Laan (2010). For an overview of the TMLE-literature, we refer to van der Laan and Rose (2011).

The TMLE-procedure follows the subsequent steps:

1. Obtain an estimator for the propensity score $P(R = 1|\boldsymbol{X})$, e.g., the MLE $\pi(\boldsymbol{X}; \hat{\boldsymbol{\psi}}_n^{\text{MLE}})$.

2. Obtain an initial estimator for the conditional mean outcome $E(Y|\boldsymbol{X})$, denoted $\hat{m}_n^{(0)}(\boldsymbol{X})$. In the simulation studies in Section 4.4, we both consider

   (a) the parametric model $m(\boldsymbol{X}; \boldsymbol{\xi})$, estimated via MLE (that is, we use $m(\boldsymbol{X}; \hat{\boldsymbol{\xi}}_n^{\text{MLE}})$ for $\hat{m}_n^{(0)}(\boldsymbol{X})$);

(b) a super-learner $\hat{m}_n^{\mathrm{SL}}(\boldsymbol{X})$. The super-learner (van der Laan et al. 2007) is a machine learning algorithm which starts from a library of estimators $\{\hat{m}_{n,j}(\boldsymbol{X}) : j = 1,\ldots,J\}$, which may consist of both parametric and nonparametric estimators. It then considers the family of all weighted averages of these estimators: $\hat{m}_{n,\boldsymbol{\omega}}(\boldsymbol{X}) = \sum_{j=1}^{J} \omega_j \hat{m}_{n,j}(\boldsymbol{X})$, $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_J)^T$, $\sum_{j=1}^{J} \omega_j = 1$ and $\omega_j \geq 0$ for $j = 1,\ldots,n$. Next choose an appropriate loss-function $\mathcal{L}_{\mathrm{SL}} : (\boldsymbol{O}, m) \to \mathcal{L}_{\mathrm{SL}}(m)(\boldsymbol{O})$ that satisfies $E(Y|\boldsymbol{X}) = \arg\min_m E\{\mathcal{L}_{\mathrm{SL}}(m)\}$, e.g., the squared error loss-function $\mathcal{L}_2(m)(\boldsymbol{O}) = R\{Y - m(\boldsymbol{X})\}^2$. The optimal weight vector $\hat{\boldsymbol{\omega}}_n$ is then defined to be the choice of $\boldsymbol{\omega}$ that minimizes the cross-validated risk with respect to this loss-function. The super-learner is defined as $\hat{m}_n^{\mathrm{SL}}(\boldsymbol{X}) = \hat{m}_{n,\hat{\boldsymbol{\omega}}_n}(\boldsymbol{X})$.

3. Fluctuation of the initial estimator: to construct an appropriate fluctuation model, we need to choose an appropriate loss-function. Gruber and van der Laan (2010) suggest to use the quasi-log-likelihood loss-function with corresponding logistic fluctuation model. The procedure now proceeds by assuming that the outcome Y is known to fall in the interval $[a,b]$ for some $a < b$. In practice, we use $a = \min_{i=1}^n Y_i - 0.1|\min_{i=1}^n Y_i|$ and $b = \max_{i=1}^n Y_i + 0.1|\max_{i=1}^n Y_i|$ for continuous outcomes, which is the default in the `tmle` R package (Gruber and van der Laan 2014). Then define the linearly transformed outcome $\widetilde{Y} = (Y-a)/(b-a)$ which falls within the unit interval $[0,1]$. If $\hat{\tilde{\mu}}_n$ is an estimator for $E(\widetilde{Y})$, $\hat{\mu}_n = (b-a)\hat{\tilde{\mu}}_n + a$ is an estimator for $E(Y)$. Without loss of generality, we assume $a = 0$ and $b = 1$ and drop the $\sim$-notation.

   Now consider the one-dimensional fluctuation model $\left\{\hat{m}_n^{(0)}(\varepsilon) : \varepsilon \in \mathbb{R}\right\}$ through the initial estimator ($\hat{m}_n^{(0)}(0) = \hat{m}_n^{(0)}$):

$$\operatorname{logit}\hat{m}_n^{(0)}(\varepsilon)(\boldsymbol{X}) = \operatorname{logit}\hat{m}_n^{(0)}(\boldsymbol{X}) + \varepsilon \widehat{H}_n(\boldsymbol{X}),$$

   with the clever covariate $\widehat{H}_n(\boldsymbol{X}) = 1/\pi(\boldsymbol{X}; \hat{\boldsymbol{\psi}}_n^{\mathrm{MLE}})$. Because the initial estimator needs to be represented as a logistic function, it needs to be bounded away from 0 and 1 and hence it is truncated at $(\zeta, 1-\zeta)$ for some small

$\zeta > 0$. Gruber and van der Laan (2010) suggest to use $\zeta = 0.005$. The optimal fluctuation is then defined by $\hat{\varepsilon}_n = \arg\min_\varepsilon n^{-1} \sum_{i=1}^n \mathcal{L}\{\hat{m}_n^{(0)}(\varepsilon)(\boldsymbol{X}_i)\}$, with $\mathcal{L}(m)(\boldsymbol{O}) = -R[Y \log\{m(\boldsymbol{X})\} + (1-Y)\log\{1-m(\boldsymbol{X})\}]$ the quasi-log-likelihood loss-function. The estimator $\hat{\varepsilon}_n$ can be obtained via a logistic regression of Y on the clever covariate $\widehat{H}_n(\boldsymbol{X})$ with the intercept $\operatorname{logit} \hat{m}_n^{(0)}$ as an offset. It follows that $\hat{\varepsilon}_n$ solves

$$\sum_{i=1}^n R_i \left\{ Y_i - \hat{m}_n^{(0)}(\hat{\varepsilon}_n)(\boldsymbol{X}_i) \right\} \widehat{H}_n(\boldsymbol{X}_i). \tag{4.17}$$

Then obtain the updated estimator $\hat{m}_n^{(1)}(\boldsymbol{X}) = \hat{m}_n^{(0)}(\hat{\varepsilon}_n)(\boldsymbol{X})$.

From this, we obtain the two TMLEs $\hat{\mu}_{n,\text{TMLE}}$ and $\hat{\mu}_{n,\text{TMLE-SL}}$ which are both calculated as $n^{-1} \sum_{i=1}^n \hat{m}_n^{(1)}(\boldsymbol{X}_i)$, with the first using $\hat{m}_n^{(0)}(\boldsymbol{X}) = m(\boldsymbol{X}; \hat{\boldsymbol{\xi}}_n^{\text{MLE}})$ and the second using $\hat{m}_n^{(0)}(\boldsymbol{X}) = \hat{m}_n^{\text{SL}}(\boldsymbol{X})$. The double robustness of both TMLEs is guaranteed by (4.17).

Both TMLEs can be fitted using the `tmle` package for R available on CRAN (Gruber and van der Laan 2014), which is used in the simulation studies in Section 4.4. The default implementation uses truncated inverse weights at $\delta = 0.025$. The super-learner can be obtained via the `SuperLearner` package for R, also available on CRAN (Polley and van der Laan 2014).

4.4 Simulation Studies

We carried out different simulation studies to compare the performance of the bias-reduced doubly robust estimator $\hat{\mu}_{n,\text{BR}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}})$ with several alternatives for the estimation of a mean outcome in the presence of incomplete data. Nuisance parameters estimated via standard MLE (e.g., solving (2.30) and (2.32)), are denoted $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{MLE}}$. We consider the standard estimators $\hat{\mu}_{n,\text{IPTW}} = n^{-1} \sum_{i=1}^n R_i Y_i / \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\text{MLE}})\}$, $\hat{\mu}_{n,\text{IMP}} = n^{-1} \sum_{i=1}^n m(\boldsymbol{X}_i; \hat{\boldsymbol{\xi}}_n^{\text{MLE}})$ and the standard doubly robust estimator $\hat{\mu}_{n,\text{MLE}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}, \hat{\boldsymbol{\xi}}_n^{\text{MLE}})$. Next, we also consider the doubly robust estimators $\hat{\mu}_{n,\text{GBM}}$ (McCaffrey et al. 2004) based on the GBM for the propensity score, the calibrated likelihood estimator $\hat{\mu}_{n,\text{TAN}}$ (Tan 2010), the projection estimator $\hat{\mu}_{n,\text{PROJ}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}, \hat{\boldsymbol{\xi}}_n^{\text{PROJ}})$ (Cao et al. 2009)

and finally the two TMLEs $\hat{\mu}_{n,\text{TMLE}}$ and $\hat{\mu}_{n,\text{TMLE-SL}}$ (van der Laan and Rose 2011), where $\hat{\mu}_{n,\text{TMLE}}$ uses ordinary least squares as an initial estimate for the conditional mean outcome, whereas $\hat{\mu}_{n,\text{TMLE-SL}}$ uses a super-learner based on a library consisting of generalized additive and linear models, random forests and adaptive polynomial splines.

For each scenario, we perform 1000 Monte Carlo runs at sample sizes of $n = 200$ and 1000. For each estimator, we calculated the Monte Carlo bias (BIAS), the root mean square error (RMSE), the median of absolute errors (MAE) and the Monte Carlo standard deviation (MCSD), that is, the empirical standard error. Occasionally, no convergence was attained for the estimator $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ at $n = 200$, as indicated below each table.

4.4.1 Scenario 1: one-covariate setting

Data-generating mechanism

The first simulation scenario considers a simple data-generating mechanism where for each i ($i = 1, \ldots, n$),

$$
\begin{aligned}
&X_i \overset{d}{=} N(0,1),\\
&R_i|X_i \overset{d}{=} \text{Ber}\{\pi_0(X_i)\} \text{ and}\\
&Y_i|X_i \overset{d}{=} N\{m_0(X_i),1\}.
\end{aligned}
$$

For each setting, the following working models are used: $\pi(X;\boldsymbol{\psi}) = \text{expit}(\psi_1 + \psi_2 X)$, $\boldsymbol{\psi} = (\psi_1,\psi_2)^T$ and $m(X;\boldsymbol{\xi}) = \xi_1 + \xi_2 X$, $\boldsymbol{\xi} = (\xi_1,\xi_2)^T$. Simulation experiments with correctly specified working models used $m_0(X) = 1 + X$ and $\pi_0(X) = \text{expit}(\gamma X)$ for $\gamma = 1,2$ (see Figure 4.3). To allow for misspecification in the outcome model, we additionally generated data using $m_0(X) = X^2$ and $\pi_0(X) = \text{expit}(\gamma X)$ for $\gamma = 1,2$. To allow for misspecification of the propensity score model, we generated data using $m_0(X) = 1 + X$ and $\pi_0(X) = \text{expit}(-4 + 1.5|X|^{0.5} + 0.75X + 0.5|X|^{1.5})$ (see Figure 4.4), as in Vansteelandt et al. (2012). Finally, we also generated data with $m_0(X) = X^2$ and $\pi_0(X) = \text{expit}(-4 + 1.5|X|^{0.5} + 0.75X + 0.5|X|^{1.5})$ to allow for misspecification of both models. In each of the settings, the target parameter

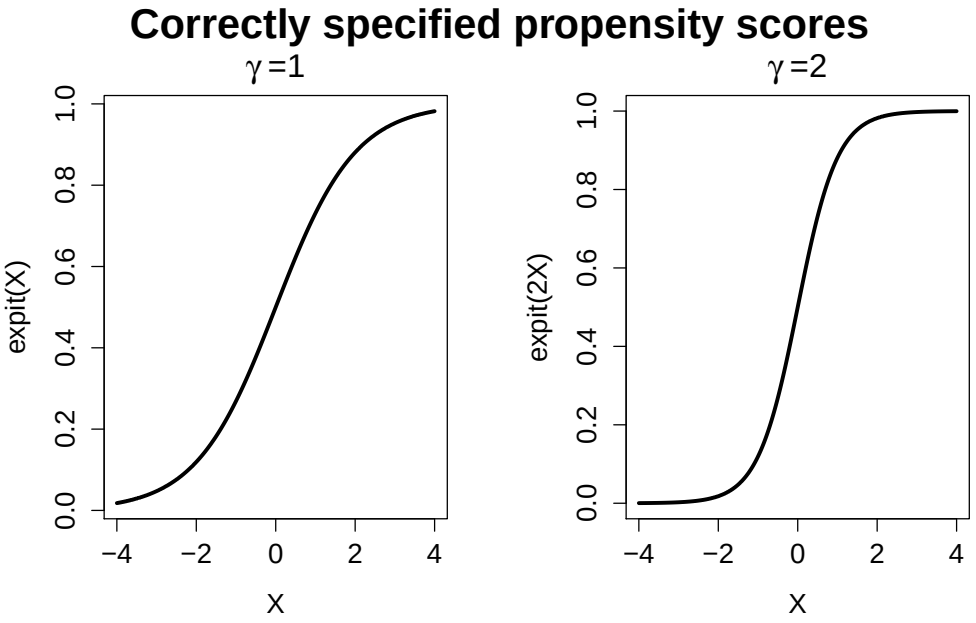


Figure 4.3: *Plot of correctly specified propensity score:* $\pi_0(X) = expit(X)$ *(left) and* $\pi_0(X) = expit(2X)$ *(right).*

$E(Y) = \mu_0$ equals one. Table 4.1 shows for each underlying propensity score, the probability $P(R = 0)$, that is, the marginal probability of the outcome Y being missing.

Table 4.1: *Marginal probability of the outcome being missing.*

PROPENSITY SCORE	$P(R=0)$
$\pi_0(X) = \text{expit}(X)$	0.50
$\pi_0(X) = \text{expit}(2X)$	0.52
$\pi_0(X) = \text{expit}(-4 + 1.5\|X\|^{0.5} + 0.75X + 0.5\|X\|^{1.5})$	0.86

Results

Results for the first simulation scenario are given in Table 4.2 ($n = 200$) and Table 4.3 ($n = 1000$). Both tables show similar results. When both working models

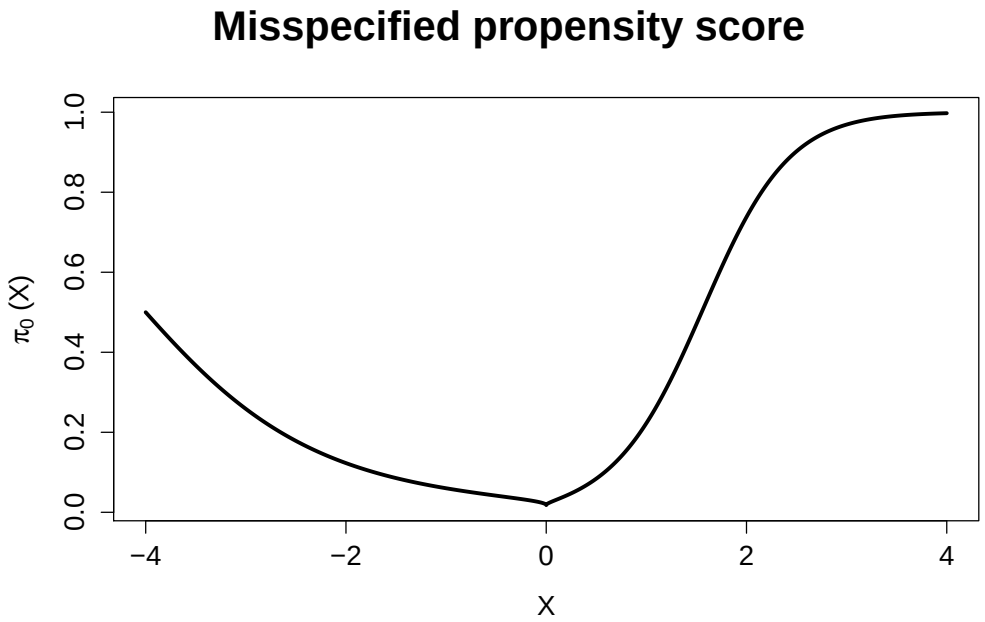


Figure 4.4: *Plot of misspecified propensity score:* $\pi_0(X) = expit(-4 + 1.5|X|^{0.5} + 0.75X + 0.5|X|^{1.5})$.

are correct and weights are not highly variable ($\gamma = 1$), all estimators perform similarly in terms of bias and precision. When at most one working model is misspecified, $\hat{\mu}_{n,\text{BR}}$ is competitive with the other doubly robust estimators in terms of RMSE. In these cases, $\hat{\mu}_{n,\text{BR}}$ shows lower or similar bias than $\hat{\mu}_{n,\text{GBM}}$, $\hat{\mu}_{n,\text{TMLE}}$ and $\hat{\mu}_{n,\text{TMLE-SL}}$, especially when the outcome model is misspecified. When the propensity score model is correctly specified, $\hat{\mu}_{n,\text{BR}}$ outperforms $\hat{\mu}_{n,\text{PROJ}}$ for the smaller sample size $n = 200$ and highly variable weights ($\gamma = 2$) and performs just slightly worse in other settings. When the propensity score working model is misspecified, $\hat{\mu}_{n,\text{BR}}$ drastically outperforms $\hat{\mu}_{n,\text{PROJ}}$. Finally, when both working models are misspecified, $\hat{\mu}_{n,\text{BR}}$ partly eliminates the bias amplification of the doubly robust estimator based on standard MLE for the nuisance parameters, although not as much as $\hat{\mu}_{n,\text{PROJ}}$ and $\hat{\mu}_{n,\text{TMLE-SL}}$. Table 4.4 suggests that this may be an artefact related to the considered data-generating mechanism and sample size: it shows that when both nuisance working models are misspecified, the bias of $\hat{\mu}_{n,\text{PROJ}}$ keeps on increasing with increasing sample size (and surprisingly also its variance) in

Table 4.2: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 1,* $n = 200$.

ESTIMATOR	BIAS	RMSE	MAE	MCSD	BIAS	RMSE	MAE	MCSD
				$n = 200$				
	OR correct, PS correct ($\gamma = 1$)				OR incorrect, PS correct ($\gamma = 1$)			
$\hat{\mu}_{n,\text{IMP}}$	−0.002	0.13	0.09	0.13	−0.349	0.40	0.35	0.19
$\hat{\mu}_{n,\text{IPTW}}$	−0.001	0.15	0.09	0.15	−0.004	0.30	0.17	0.30
$\hat{\mu}_{n,\text{MLE}}$	−0.001	0.13	0.09	0.13	−0.018	0.33	0.19	0.33
$\hat{\mu}_{n,\text{BR}}$	−0.001	0.13	0.09	0.13	−0.030	0.21	0.14	0.20
$\hat{\mu}_{n,\text{TAN}}$	−0.001	0.13	0.09	0.13	−0.032	0.18	0.13	0.18
$\hat{\mu}_{n,\text{PROJ}}$	−0.002	0.15	0.10	0.15	−0.019	0.17	0.12	0.17
$\hat{\mu}_{n,\text{GBM}}$	−0.001	0.13	0.09	0.13	−0.190	0.27	0.20	0.20
$\hat{\mu}_{n,\text{TMLE}}$	−0.001	0.13	0.09	0.13	−0.038	0.27	0.18	0.26
$\hat{\mu}_{n,\text{TMLE-SL}}$	0.000	0.13	0.09	0.13	−0.027	0.17	0.12	0.17
	OR correct, PS correct ($\gamma = 2$)				OR incorrect, PS correct ($\gamma = 2$)			
$\hat{\mu}_{n,\text{IMP}}$	−0.002	0.14	0.09	0.14	−0.81	0.84	0.81	0.22
$\hat{\mu}_{n,\text{IPTW}}$	0.003	0.26	0.12	0.26	−0.01	0.92	0.28	0.92
$\hat{\mu}_{n,\text{MLE}}$	−0.002	0.20	0.12	0.20	−0.06	1.19	0.44	1.19
$\hat{\mu}_{n,\text{BR}}$	−0.001	0.19	0.12	0.19	−0.11	0.26	0.17	0.24
$\hat{\mu}_{n,\text{TAN}}$	−0.001	0.21	0.13	0.21	−0.10	0.25	0.16	0.23
$\hat{\mu}_{n,\text{PROJ}}$	−0.002	0.33	0.18	0.33	−0.07	0.34	0.19	0.33
$\hat{\mu}_{n,\text{GBM}}$	−0.001	0.15	0.10	0.15	−0.54	0.60	0.53	0.26
$\hat{\mu}_{n,\text{TMLE}}$	0.003	0.17	0.11	0.17	−0.15	0.36	0.26	0.33
$\hat{\mu}_{n,\text{TMLE-SL}}$	0.008	0.18	0.12	0.18	−0.08	0.24	0.15	0.23
	OR correct, PS incorrect				OR incorrect, PS incorrect			
$\hat{\mu}_{n,\text{IMP}}$	−0.001	0.27	0.17	0.27	0.54	0.96	0.72	0.80
$\hat{\mu}_{n,\text{IPTW}}$	−1.612	5.32	0.95	5.07	6.13	16.53	3.07	15.36
$\hat{\mu}_{n,\text{MLE}}$	−0.008	1.04	0.31	1.04	5.50	11.46	2.99	10.06
$\hat{\mu}_{n,\text{BR}}$	−0.003	0.29	0.18	0.29	1.03	1.23	1.05	0.68
$\hat{\mu}_{n,\text{TAN}}$	−0.011	0.30	0.19	0.30	0.43	0.67	0.47	0.51
$\hat{\mu}_{n,\text{PROJ}}$	−0.028	0.57	0.23	0.57	−0.09	0.62	0.24	0.62
$\hat{\mu}_{n,\text{GBM}}$	−0.004	0.28	0.19	0.28	0.33	0.71	0.48	0.63
$\hat{\mu}_{n,\text{TMLE}}$	−0.006	0.28	0.18	0.28	1.10	1.31	1.14	0.70
$\hat{\mu}_{n,\text{TMLE-SL}}$	−0.012	0.28	0.19	0.28	0.32	0.52	0.38	0.41

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MAE, median of absolute errors; MCSD, Monte Carlo standard deviation; OR, outcome regression; PS, propensity score. No convergence for $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ was attained in five of the 1000 runs for the settings OR correct, PS correct ($\gamma = 2$) and OR incorrect, PS correct ($\gamma = 2$) and in three of the 1000 runs for the settings OR correct, PS incorrect and OR incorrect, PS incorrect.

Table 4.3: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 1,* $n = 1000$.

ESTIMATOR	BIAS	RMSE	MAE	MCSD	BIAS	RMSE	MAE	MCSD
				$n = 1000$				
	OR correct, PS correct ($\gamma = 1$)				OR incorrect, PS correct ($\gamma = 1$)			
$\hat{\mu}_{n,\text{OR}}$	0.0037	0.057	0.039	0.057	−0.349	0.36	0.35	0.09
$\hat{\mu}_{n,\text{IPTW}}$	0.0022	0.064	0.044	0.064	0.006	0.13	0.08	0.13
$\hat{\mu}_{n,\text{MLE}}$	0.0029	0.059	0.039	0.058	0.003	0.15	0.10	0.15
$\hat{\mu}_{n,\text{BR}}$	0.0031	0.059	0.039	0.058	−0.003	0.09	0.07	0.09
$\hat{\mu}_{n,\text{TAN}}$	0.0033	0.059	0.038	0.059	−0.005	0.08	0.06	0.08
$\hat{\mu}_{n,\text{PROJ}}$	0.0038	0.060	0.039	0.060	−0.002	0.07	0.05	0.07
$\hat{\mu}_{n,\text{GBM}}$	0.0032	0.058	0.040	0.058	−0.138	0.16	0.14	0.08
$\hat{\mu}_{n,\text{TMLE}}$	0.0030	0.058	0.039	0.058	−0.004	0.12	0.09	0.12
$\hat{\mu}_{n,\text{TMLE-SL}}$	0.0031	0.058	0.039	0.058	−0.002	0.07	0.05	0.07
	OR correct, PS correct ($\gamma = 2$)				OR incorrect, PS correct ($\gamma = 2$)			
$\hat{\mu}_{n,\text{OR}}$	0.0033	0.065	0.044	0.065	−0.810	0.82	0.81	0.10
$\hat{\mu}_{n,\text{IPTW}}$	0.0051	0.120	0.066	0.120	−0.023	0.34	0.17	0.34
$\hat{\mu}_{n,\text{MLE}}$	−0.0003	0.090	0.057	0.090	−0.041	0.50	0.25	0.49
$\hat{\mu}_{n,\text{BR}}$	0.0004	0.085	0.057	0.085	−0.052	0.14	0.09	0.13
$\hat{\mu}_{n,\text{TAN}}$	0.0005	0.091	0.060	0.091	−0.042	0.11	0.07	0.10
$\hat{\mu}_{n,\text{PROJ}}$	−0.0014	0.109	0.069	0.109	−0.027	0.12	0.07	0.11
$\hat{\mu}_{n,\text{GBM}}$	0.0026	0.072	0.048	0.071	−0.407	0.43	0.40	0.12
$\hat{\mu}_{n,\text{TMLE}}$	0.0011	0.080	0.054	0.080	−0.124	0.19	0.14	0.15
$\hat{\mu}_{n,\text{TMLE-SL}}$	0.0011	0.081	0.055	0.081	−0.041	0.11	0.07	0.10
	OR correct, PS incorrect				OR incorrect, PS incorrect			
$\hat{\mu}_{n,\text{OR}}$	−0.0056	0.11	0.07	0.11	0.70	0.78	0.70	0.35
$\hat{\mu}_{n,\text{IPTW}}$	−1.8704	2.50	1.46	1.65	7.34	10.03	5.59	6.84
$\hat{\mu}_{n,\text{MLE}}$	−0.0264	0.52	0.20	0.52	7.24	9.64	5.57	6.37
$\hat{\mu}_{n,\text{BR}}$	−0.0057	0.11	0.08	0.11	1.24	1.28	1.22	0.33
$\hat{\mu}_{n,\text{TAN}}$	−0.0043	0.12	0.08	0.12	0.43	0.47	0.42	0.20
$\hat{\mu}_{n,\text{PROJ}}$	−0.0007	0.50	0.17	0.50	0.09	0.57	0.19	0.56
$\hat{\mu}_{n,\text{GBM}}$	−0.0002	0.13	0.09	0.13	0.20	0.27	0.21	0.18
$\hat{\mu}_{n,\text{TMLE}}$	−0.0052	0.12	0.08	0.12	1.20	1.24	1.20	0.31
$\hat{\mu}_{n,\text{TMLE-SL}}$	−0.0071	0.12	0.08	0.12	0.10	0.18	0.12	0.15

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MAE, median of absolute errors; MCSD, Monte Carlo standard deviation; OR, outcome regression; PS, propensity score.

contrast to the bias of $\hat{\mu}_{n,\text{BR}}$ which remains stable; this is in line with the fact that $\hat{\mu}_{n,\text{BR}}$ minimizes the asymptotic bias. Also note that the standard deviation of the bias-reduced doubly robust estimator decreases at root-n rate, which is not always

the case for the other estimators (see also Table 4.8). Finally, the smaller bias of $\hat{\mu}_{n,\text{TMLE-SL}}$ and $\hat{\mu}_{n,\text{TAN}}$ under misspecification of both working models is not unexpected because of the richer working models for the conditional mean outcome on which these rely.

Table 4.4: *Monte Carlo bias and standard deviation based on* 1000 *Monte Carlo replications for the bias-reduced estimation strategy as compared to standard MLE and the projection estimator for Scenario 1 when both working models are misspecified.*

		ESTIMATOR		
		$\hat{\mu}_{n,\text{MLE}}$	$\hat{\mu}_{n,\text{BR}}$	$\hat{\mu}_{n,\text{PROJ}}$
$n = 1000$	BIAS	7.24	1.24	0.09
	MCSD	6.37	2.63	1.88
$n = 5000$	BIAS	7.20	1.29	0.28
	MCSD	2.63	0.15	0.41
$n = 10000$	BIAS	7.28	1.30	0.42
	MCSD	1.88	0.11	0.40
$n = 50000$	BIAS	7.27	1.30	0.66
	MCSD	0.84	0.05	0.38
$n = 100000$	BIAS	7.25	1.30	0.76
	MCSD	0.58	0.03	0.36
$n = 500000$	BIAS	7.26	1.30	0.90
	MCSD	0.270	0.015	0.320
$n = 1000000$	BIAS	7.27	1.30	0.96
	MCSD	0.180	0.011	0.290

NOTE: BIAS, Monte Carlo Bias; MCSD, Monte Carlo standard deviation.

Table 4.5 shows the performance of the sandwich estimator for the standard error for $\hat{\mu}_{n,\text{BR}}$ computed as the empirical variance of (3.5) and confirms the asymptotic result of Corollary 4.1. Not surprisingly, there is under-coverage of the 95% confidence intervals when the inverse propensity score becomes extreme, especially at the smaller sample size of $n = 200$. When weights are extreme, convergence to the normal limit distribution happens more slowly. The coverage is better at $n = 1000$.

Table 4.5: *Performance of standard error estimates and confidence intervals for the bias-reduced estimation strategy based on* 1000 *Monte Carlo replications in Scenario 1.*

	$n = 200$			$n = 1000$		
SETTING	MCSD	ASSE	COV	MCSD	ASSE	COV
OR correct, PS correct ($\gamma = 1$)	0.13	0.13	0.94	0.06	0.06	0.96
OR correct, PS correct ($\gamma = 2$)	0.19	0.15	0.88	0.09	0.08	0.91
OR incorrect, PS correct ($\gamma = 1$)	0.20	0.17	0.88	0.09	0.09	0.94
OR incorrect, PS correct ($\gamma = 2$)	0.24	0.17	0.77	0.13	0.10	0.80
OR correct, PS incorrect	0.29	0.22	0.85	0.11	0.11	0.94
OR incorrect, PS incorrect	0.68	0.43	0.34	0.33	0.28	0.01

NOTE: MCSD, Monte Carlo standard deviation; ASSE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; OR, outcome regression; PS, propensity score.

4.4.2 Scenario 2: Kang and Schafer setting

Data-generating mechanism

The second simulation scenario is taken from Kang and Schafer (2007a). For each i $(i = 1, \ldots, n)$,

$$\begin{aligned} \boldsymbol{Z}_i &\stackrel{d}{=} \mathbf{N}(\mathbf{0}, \mathbf{I}), \\ R_i | \boldsymbol{Z}_i &\stackrel{d}{=} \text{Ber}\{\pi_0(\boldsymbol{Z}_i)\} \text{ and} \\ Y_i | \boldsymbol{Z}_i &\stackrel{d}{=} N\{m_0(\boldsymbol{Z}_i), 1\}, \end{aligned}$$

where $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$, $\mathbf{I}$ is the 4×4 identity matrix, $\pi_0(\boldsymbol{Z}) = \text{expit}(-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4)$ and $m_0(\boldsymbol{Z}) = 210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4$. Misspecified working models are linear for the outcome model and logistic for the propensity score model, with covariates $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$, with

$$\begin{aligned} X_1 &= \exp(Z_1/2), \\ X_2 &= Z_2/\{1 + \exp(Z_1)\} + 10, \\ X_3 &= (Z_1 Z_3/25 + 0.6)^3 \text{ and} \end{aligned}$$

$$X_4 = (Z_2 + Z_4 + 20)^2.$$

The target parameter $E(Y) = \mu_0$ equals 210. In all cases, the marginal probability $P(R = 0)$ of the outcome Y being missing equals 0.5. We limit ourselves to the realistic settings where the working models both use either the covariates Z_k or the covariates X_k ($k = 1, \ldots, 4$) and thus both working models are correctly specified or both working models are incorrectly specified.

Results

Table 4.6 shows the simulation results for two scenarios where either $R = 1$ or $R = 0$ denotes the data that are observed for $n = 200$ and Table 4.7 shows the results for $n = 1000$. As the theory dictates, all doubly robust estimators show similar behavior when both working models are correctly specified. When the observed outcome RY is used, $\hat{\mu}_{n,\text{MLE}}$ shows severe erratic behavior, corresponding to the results of Kang and Schafer (2007a), but this behavior is partially eliminated when using $(1-R)Y$ as the observed outcome, in which case $\hat{\mu}_{n,\text{MLE}}$ now outperforms $\hat{\mu}_{n,\text{OR}}$ (Robins et al. 2007). The DR estimator $\hat{\mu}_{n,\text{BR}}$ does not show this severe erratic behavior for both RY and $(1-R)Y$. In line with Theorem 4.1, it has smaller bias compared to standard MLE. There is no single alternative outperforming the others for both sample sizes and both settings RY and $(1-R)Y$. Overall, $\hat{\mu}_{n,\text{BR}}$ shows competitive performance with the other doubly robust estimators (see also Table 4.8 for additional results with increasing sample sizes).

Table 4.9 shows the performance of the sandwich estimator for the standard error of $\hat{\mu}_{n,\text{BR}}$ computed as the empirical variance of (3.5) and confirms the asymptotic result of Corollary 4.1.

4.4.3 Conclusion

The simulations studies show that all alternative estimation strategies outperform the standard procedures $\hat{\mu}_{n,\text{OR}}$, $\hat{\mu}_{n,\text{IPTW}}$ and $\hat{\mu}_{n,\text{MLE}}$ to estimate the mean outcome with ignorable missingness under misspecification of one or two working models. It is seen that when at most one working model is misspecified, the bias-reduced doubly

Table 4.6: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 2,* $n = 200$.

ESTIMATOR	BIAS	RMSE	MAE	MCSD	BIAS	RMSE	MAE	MCSD
	observed outcome RY				observed outcome $(1-R)Y$			
				$n = 200$				
	OR correct, PS correct				OR correct, PS correct			
$\hat{\mu}_{n,\text{OR}}$	0.092	2.52	1.68	2.52	0.088	2.53	1.68	2.53
$\hat{\mu}_{n,\text{IPTW}}$	−1.761	28.15	13.22	28.10	−0.254	16.64	8.22	16.65
$\hat{\mu}_{n,\text{MLE}}$	0.099	2.53	1.70	2.53	0.085	2.53	1.73	2.53
$\hat{\mu}_{n,\text{BR}}$	0.090	2.54	1.71	2.54	0.095	2.54	1.69	2.54
$\hat{\mu}_{n,\text{TAN}}$	0.094	2.53	1.72	2.53	0.085	2.53	1.69	2.53
$\hat{\mu}_{n,\text{PROJ}}$	0.090	2.55	1.71	2.55	0.079	2.54	1.72	2.54
$\hat{\mu}_{n,\text{GBM}}$	0.093	2.53	1.70	2.53	0.088	2.53	1.67	2.53
$\hat{\mu}_{n,\text{TMLE}}$	0.032	2.53	1.72	2.53	0.238	2.55	1.77	2.54
$\hat{\mu}_{n,\text{TMLE-SL}}$	0.031	2.53	1.71	2.53	0.241	2.55	1.78	2.54
	OR incorrect, PS incorrect				OR incorrect, PS incorrect			
$\hat{\mu}_{n,\text{OR}}$	−0.17	3.60	2.51	3.59	7.15	7.76	7.21	3.01
$\hat{\mu}_{n,\text{IPTW}}$	68.77	453.60	18.43	448.58	-0.80	12.18	6.24	12.16
$\hat{\mu}_{n,\text{MLE}}$	−15.15	88.60	4.40	87.34	4.76	6.05	5.03	3.74
$\hat{\mu}_{n,\text{BR}}$	−2.24	4.45	2.78	3.85	3.44	4.63	3.55	3.10
$\hat{\mu}_{n,\text{TAN}}$	−2.55	4.31	2.96	3.47	4.76	5.79	4.85	3.30
$\hat{\mu}_{n,\text{PROJ}}$	−0.04	3.93	2.58	3.93	1.00	3.60	2.36	3.46
$\hat{\mu}_{n,\text{GBM}}$	−0.22	3.46	2.42	3.45	5.84	6.58	5.86	3.03
$\hat{\mu}_{n,\text{TMLE}}$	−4.38	6.20	4.08	4.39	4.43	5.56	4.47	3.35
$\hat{\mu}_{n,\text{TMLE-SL}}$	−2.31	4.02	2.72	3.29	3.75	5.07	3.95	3.41

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MAE, median of absolute errors; MCSD, Monte Carlo standard deviation; OR, outcome regression; PS, propensity score. No convergence for $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ was attained in 13 of the 1000 runs for the settings OR correct, PS correct, $n = 200$ for both the observed outcome RY and $(1-R)Y$ and in five of the 1000 runs for the setting OR incorrect, PS incorrect, $n = 200$ for the observed outcome RY.

robust estimation principle performs well in all different settings and particularly well when weights become highly variable. Moreover, the quality of a given nuisance parameter estimation strategy depends on the true underlying data-generating mechanism in finite sample sizes and no alternative appears uniformly superior to all others. In Scenario 1, when both working models are misspecified, $\hat{\mu}_{n,\text{TMLE-SL}}$ outperforms the others, especially for large sample size, which is a result of the richer class of working models on which the estimator relies. In Scenario 2, there is no single alternative outperforming the others for both sample sizes and both

Table 4.7: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 2,* $n = 1000$.

ESTIMATOR	BIAS	RMSE	MAE	MCSD	BIAS	RMSE	MAE	MCSD
	observed outcome RY				observed outcome $(1-R)Y$			
	$n = 1000$							
	OR correct, PS correct				OR correct, PS correct			
$\hat{\mu}_{n,\text{OR}}$	0.023	1.12	0.76	1.12	0.024	1.13	0.76	1.13
$\hat{\mu}_{n,\text{IPTW}}$	−0.282	11.27	6.61	11.27	0.052	7.67	3.93	7.67
$\hat{\mu}_{n,\text{MLE}}$	0.023	1.12	0.76	1.12	0.022	1.13	0.78	1.13
$\hat{\mu}_{n,\text{BR}}$	0.022	1.12	0.76	1.12	0.024	1.13	0.78	1.13
$\hat{\mu}_{n,\text{TAN}}$	0.021	1.12	0.76	1.12	0.024	1.13	0.77	1.13
$\hat{\mu}_{n,\text{PROJ}}$	0.024	1.12	0.76	1.12	0.025	1.13	0.77	1.13
$\hat{\mu}_{n,\text{GBM}}$	0.022	1.12	0.76	1.12	0.024	1.13	0.77	1.13
$\hat{\mu}_{n,\text{TMLE}}$	0.013	1.12	0.76	1.12	0.061	1.13	0.76	1.13
$\hat{\mu}_{n,\text{TMLE-SL}}$	0.013	1.12	0.76	1.12	0.060	1.13	0.77	1.13
	OR incorrect, PS incorrect				OR incorrect, PS incorrect			
$\hat{\mu}_{n,\text{OR}}$	−0.46	1.64	1.11	1.58	7.18	7.31	7.14	1.35
$\hat{\mu}_{n,\text{IPTW}}$	161.92	1194.95	36.92	1184.52	−1.00	4.93	3.12	4.83
$\hat{\mu}_{n,\text{MLE}}$	−53.71	469.04	8.81	466.19	4.51	4.83	4.51	1.73
$\hat{\mu}_{n,\text{BR}}$	−3.21	3.63	3.10	1.69	2.97	3.25	2.91	1.34
$\hat{\mu}_{n,\text{TAN}}$	−2.92	3.29	3.02	1.51	4.35	4.62	4.28	1.55
$\hat{\mu}_{n,\text{PROJ}}$	−1.55	2.03	1.56	1.32	1.39	1.92	1.53	1.33
$\hat{\mu}_{n,\text{GBM}}$	−0.60	1.56	1.09	1.44	4.83	5.03	4.75	1.39
$\hat{\mu}_{n,\text{TMLE}}$	−4.73	5.13	4.73	1.99	4.16	4.43	4.11	1.52
$\hat{\mu}_{n,\text{TMLE-SL}}$	−2.25	2.76	2.34	1.61	2.65	3.12	2.57	1.65

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MAE, median of absolute errors; MCSD, Monte Carlo standard deviation; OR, outcome regression; PS, propensity score.

settings RY and $(1-R)Y$. Overall, we believe that the bias-reduced doubly robust estimator is particularly useful in small to moderate samples (of the order used in the simulation experiments considered in this section) and in settings where the weights obtained via MLE are highly variable.

Table 4.8: *Monte Carlo bias and standard deviation based on* 1000 *Monte Carlo replications for the bias-reduced estimation strategy as compared to standard MLE and the projection estimator for Scenario 2 when both working models are misspecified.*

		ESTIMATOR			ESTIMATOR		
		$\hat{\mu}_{n,\text{MLE}}$	$\hat{\mu}_{n,\text{BR}}$	$\hat{\mu}_{n,\text{PROJ}}$	$\hat{\mu}_{n,\text{MLE}}$	$\hat{\mu}_{n,\text{BR}}$	$\hat{\mu}_{n,\text{PROJ}}$
		Scenario 2, RY			Scenario 2, $(1-R)Y$		
$n = 1000$	BIAS	−53.71	−3.21	−1.55	4.51	2.97	1.39
	MCSD	466.19	1.69	1.32	1.73	1.34	1.33
$n = 5000$	BIAS	−202.18	−3.86	−2.09	4.50	2.85	1.67
	MCSD	1911.11	1.12	0.64	0.77	0.61	1.67
$n = 10000$	BIAS	−1649.86	−4.14	−2.24	4.45	2.80	1.74
	MCSD	30926.76	0.97	0.60	0.54	0.44	0.46
$n = 50000$	BIAS	−1334.08	−4.57	−2.55	4.48	2.80	1.84
	MCSD	23987.58	0.69	0.94	0.23	0.19	0.22
$n = 100000$	BIAS	−1702.68	−4.73	−2.77	4.47	2.79	1.87
	MCSD	19376.55	0.0.64	1.08	0.16	0.13	0.16
$n = 500000$	BIAS	−52240.56	−5.04	−3.49	4.47	2.79	1.91
	MCSD	1082593.64	0.62	2.51	0.08	0.06	0.04
$n = 1000000$	BIAS	−15431.40	−5.15	−3.89	4.47	2.79	1.93
	MCSD	218954.30	0.55	3.31	0.06	0.04	0.08

NOTE: BIAS, Monte Carlo Bias; MCSD, Monte Carlo standard deviation.

Table 4.9: *Performance of standard error estimates and confidence intervals for the bias-reduced estimation strategy based on* 1000 *Monte Carlo replications in Scenario 2.*

	$n = 200$			$n = 1000$		
SETTING	MCSD	ASSE	COV	MCSD	ASSE	COV
OR correct, PS correct (RY)	2.54	2.57	0.96	1.12	1.15	0.96
OR correct, PS correct ($(1-R)Y$)	2.54	2.57	0.95	1.13	1.15	0.96
OR incorrect, PS incorrect (RY)	3.85	2.95	0.82	1.69	1.34	0.38
OR incorrect, PS incorrect ($(1-R)Y$)	3.10	2.73	0.73	1.34	1.25	0.35

NOTE: MCSD, Monte Carlo standard deviation; ASSE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; OR, outcome regression; PS, propensity score.

4.5 Extension to Other Doubly and Multiply Robust Estimators

In this section, we give several examples to illustrate the broad applicability of the bias-reduced doubly robust estimation principle.

4.5.1 Marginal treatment effects

Marginal treatment effect for a dichotomous treatment

Consider i.i.d. data $\{\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$, where Y_i is the outcome of interest, A_i is a dichotomous treatment taking values zero and one and $\boldsymbol{X}_i$ is a sufficient set of covariates to control for confounding of the treatment effect, in the sense that $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$ for $a \in \{0,1\}$ (no-unmeasured confounders assumption). Here, $Y(a)$ denotes the counterfactual outcome for treatment level $a \in \{0,1\}$, which is linked to the observed data through the consistency assumption (i.e., $Y(a) = Y$ iff $A = a$).

To obtain a doubly robust estimator for the marginal treatment effect $\tau = E\{Y(1)\} - E\{Y(0)\} = \mu_0^{(1)} - \mu_0^{(0)}$, we need three working models: a model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for the propensity score $P(A = 1|\boldsymbol{X})$ (for which we assume positivity: $1 > 1 - \delta \geq P(A = 1|\boldsymbol{X}) \geq \delta > 0$ with probability one, see van der Laan and Rose (2011), Chapter 10) and models $m^{(a)}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)})$ for the conditional mean outcome $E(Y|A = a, \boldsymbol{X})$ for $a \in \{0,1\}$. We estimate the treatment effect as $\hat{\tau}_n = \hat{E}_n\{Y(1)\} - \hat{E}_n\{Y(0)\} = \hat{\mu}_{n,\mathrm{DR}}^{(1)} - \hat{\mu}_{n,\mathrm{DR}}^{(0)}$ where a doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}^{(a)} \equiv \hat{\mu}_{n,\mathrm{DR}}^{(a)}(\boldsymbol{\psi}, \boldsymbol{\alpha}^{(a)})$ of $\mu_0^{(a)}$ is obtained as the solution to the estimating equation

$$\sum_{i=1}^{n} \phi^{(a)}(\boldsymbol{O}_i; \hat{\mu}_n^{(a)}, \boldsymbol{\psi}, \boldsymbol{\alpha}^{(a)})\} = 0$$

for $a \in \{0,1\}$ (Scharfstein et al. 1999b), where

$$\begin{aligned}
\phi^{(1)}(\boldsymbol{O};\mu^{(1)},\boldsymbol{\psi},\boldsymbol{\alpha}^{(1)}) &= \frac{AY}{\pi(\boldsymbol{X};\boldsymbol{\psi})} - \frac{A-\pi(\boldsymbol{X};\boldsymbol{\psi})}{\pi(\boldsymbol{X};\boldsymbol{\psi})} m^{(1)}(\boldsymbol{X};\boldsymbol{\alpha}^{(1)}) - \mu^{(1)}, \\
\phi^{(0)}(\boldsymbol{O};\mu^{(0)},\boldsymbol{\psi},\boldsymbol{\alpha}^{(0)}) &= \frac{(1-A)Y}{1-\pi(\boldsymbol{X};\boldsymbol{\psi})} + \frac{A-\pi(\boldsymbol{X};\boldsymbol{\psi})}{1-\pi(\boldsymbol{X};\boldsymbol{\psi})} m^{(0)}(\boldsymbol{X};\boldsymbol{\alpha}^{(0)}) - \mu^{(0)}.
\end{aligned}$$

The proposed estimation strategy proceeds by setting the gradients w.r.t. the nuisance parameters equal to zero, which amounts to solving $(\hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR}}, \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR}})$ from the system

$$\mathbf{0} = \sum_{i=1}^{n} \left[\left\{ 1 - \frac{A_i}{\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR}})} \right\} m^{(1)}_{\boldsymbol{\alpha}^{(1)}}(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR}}) \right], \tag{4.18}$$

$$\mathbf{0} = \sum_{i=1}^{n} \left[A_i \left\{ Y_i - m^{(1)}(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR}}) \right\} \frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR}})}{\pi^2(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR}})} \right] \tag{4.19}$$

and solving $(\hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR}}, \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR}})$ from the system

$$\mathbf{0} = \sum_{i=1}^{n} \left[\left\{ 1 - \frac{1 - A_i}{1 - \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR}})} \right\} m^{(0)}_{\boldsymbol{\alpha}^{(0)}}(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR}}) \right], \tag{4.20}$$

$$\mathbf{0} = \sum_{i=1}^{n} \left[(1 - A_i) \left\{ Y_i - m^{(0)}(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR}}) \right\} \frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR}})}{\{1 - \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR}})\}^2} \right], \tag{4.21}$$

where $\pi_{\boldsymbol{\psi}}(\boldsymbol{X}; \boldsymbol{\psi}) = \partial \pi(\boldsymbol{X}; \boldsymbol{\psi}) / \partial \boldsymbol{\psi}$ and $m^{(a)}_{\boldsymbol{\alpha}^{(a)}}(\boldsymbol{X}; \boldsymbol{\alpha}^{(a)}) = \partial m^{(a)}(\boldsymbol{X}; \boldsymbol{\alpha}^{(a)}) / \partial \boldsymbol{\alpha}^{(a)}$, for $a \in \{0, 1\}$. This results in estimators with similar properties as the estimators constructed in Section 4.3. Note that the doubly robust estimators $\hat{\mu}^{(1)}_{n,\text{DR}}$ and $\hat{\mu}^{(0)}_{n,\text{DR}}$, while relying on the same working model $\pi(\boldsymbol{X}; \boldsymbol{\psi})$ for the propensity score, use different estimators for the nuisance parameter indexing that model. In particular, (4.19) forces the model to fit well in covariate regions with low propensity score, where not being treated is more likely whereas (4.21) forces the model to fit well in covariate regions with high propensity score, where being treated is more likely. Additionally, (4.18) ensures stability of inverse weights equalling one over the propensity score, while (4.20) ensures stability of inverse weights equalling one over one minus the propensity score. This illustrates that the nuisance parameter estimators adapt to the considered estimand. The use of these estimators is illustrated in a re-analysis of the SUPPORT study in Section 4.6.

Marginal Structural Models (MSMs) for point-treatment data

A more general development works under the marginal structural model (Robins et al. 2000)

$$E\{Y(a)\} = \beta_0 + \beta_1 a,$$

where $a \in \mathcal{A}$ may be a continuous exposure level with $\mathcal{A}$ the support of A. This model is structural because it models part of a counterfactual-distribution and implies no restrictions on the observed data distribution. It is marginal because it models the marginal mean of the counterfactual $Y(a)$.

Let $f(a)$ be a user-specified density function for A. Then consider the doubly robust estimators obtained by solving the following estimating equations, assuming $\sup_{a\in\mathcal{A}} f(a)/f_0(a|\boldsymbol{X}) < \infty$ with probability one with $f_0(a|\boldsymbol{X})$ the true conditional density function of the treatment given covariates (Robins 1999b; van der Laan and Robins 2003, sec. 6.3):

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n} \begin{bmatrix} U_{\beta_0}(\boldsymbol{O}_i;\beta_0,\beta_1,\boldsymbol{\psi},\boldsymbol{\alpha}) \\ U_{\beta_1}(\boldsymbol{O}_i;\beta_0,\beta_1,\boldsymbol{\psi},\boldsymbol{\alpha}) \end{bmatrix} \\ &= \sum_{i=1}^{n} \Bigg[\boldsymbol{d}(A_i) \frac{f(A_i)}{f(A_i|\boldsymbol{X}_i;\boldsymbol{\psi})} \{Y_i - m(A_i,\boldsymbol{X}_i;\boldsymbol{\alpha})\} \\ &\qquad + \int_{\mathcal{A}} \boldsymbol{d}(a) \{m(a,\boldsymbol{X}_i;\boldsymbol{\alpha}) - \beta_0 - \beta_1 a\} f(a) \mathrm{d}a \Bigg],\end{aligned}$$

where $\boldsymbol{d}(a) = \mathrm{var}_f^{-1}(A)(1,a)^T$, $f(A|\boldsymbol{X};\boldsymbol{\psi})$ is a working model ($\mathcal{M}(\boldsymbol{\psi})$) for the conditional density function of the treatment given covariates, $m(A,\boldsymbol{X};\boldsymbol{\alpha})$ is a working model ($\mathcal{M}(\boldsymbol{\alpha})$) for the conditional mean $E(Y|A,\boldsymbol{X})$ and $(\boldsymbol{\psi}^T,\boldsymbol{\alpha}^T)^T$ are unknown finite-dimensional nuisance parameters.

With interest in the MSM-parameter β_1 (parameterizing the causal effect of a unit increase in the exposure: $\beta_1 = E\{Y(a+1)\} - E\{Y(a)\}$), we focus on the doubly robust influence function for β_1 considering fixed nuisance parameter values, but taking into account that the other MSM-parameter β_0 is unknown (i.e., $\phi = U_{\beta_1} - E(\partial U_{\beta_1}/\partial\beta_0)E^{-1}(\partial U_{\beta_0}/\partial\beta_0)U_{\beta_0}$),

$$\phi(\boldsymbol{O};\beta_1,\boldsymbol{\psi},\boldsymbol{\alpha})$$

$$= \mathrm{var}_f^{-1}(A)\left\{A - E_f(A)\right\}\frac{f(A)}{f(A|\boldsymbol{X};\boldsymbol{\psi})}\left\{Y - m(A,\boldsymbol{X};\boldsymbol{\alpha})\right\}$$
$$+ \mathrm{var}_f^{-1}(A)\int_{\mathcal{A}}\left\{a - E_f(A)\right\}m(a,\boldsymbol{X};\boldsymbol{\alpha})f(a)\mathrm{d}a - \beta_1,$$

where $E_f(A)$ and $\mathrm{var}_f(A)$ are evaluated according to the density $f(a)$, e.g., under the empirical mean $n^{-1}\sum_{i=1}^n f(a|\boldsymbol{X}_i;\boldsymbol{\psi})\}$. The resulting doubly robust estimator is denoted $\hat{\beta}_{n,1,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\alpha})$ and can be obtained as

4

$$\hat{\beta}_{n,1,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\alpha})$$
$$= \mathrm{var}_f^{-1}(A)\left[n^{-1}\sum_{i=1}^n\left\{A_i - E_f(A)\right\}\frac{f(A_i)}{f(A_i|\boldsymbol{X}_i;\boldsymbol{\psi})}\left\{Y_i - m(A_i,\boldsymbol{X}_i;\boldsymbol{\alpha})\right\}\right.$$
$$\left. + \int_{\mathcal{A}}\left\{a - E_f(A)\right\}m(a,\boldsymbol{X}_i;\boldsymbol{\alpha})f(a)\mathrm{d}a\right].$$

The gradients of $\phi(\boldsymbol{O};\beta_1,\boldsymbol{\psi},\boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ then define the estimating functions for $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$; that is, solve $(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})$ from the system

$$\mathbf{0} = \sum_{i=1}^n\int_{\mathcal{A}}\{a - E_f(A)\}\left[1 - \frac{I(A_i = a)}{f(a|\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}\right]$$
$$\times m_{\boldsymbol{\alpha}}(a,\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})f(a)\mathrm{d}a, \tag{4.22}$$
$$\mathbf{0} = \sum_{i=1}^n\{A_i - E_f(A)\}\frac{f(A_i)}{f(A_i|\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}S_{\boldsymbol{\psi}}(A_i|\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})$$
$$\times\{Y_i - m(A_i,\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})\}, \tag{4.23}$$

where $m_{\boldsymbol{\alpha}}(A,\boldsymbol{X};\boldsymbol{\alpha}) = \partial m(A,\boldsymbol{X};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ and $S_{\boldsymbol{\psi}}(A|\boldsymbol{X};\boldsymbol{\psi})$ equals the score w.r.t. $\boldsymbol{\psi}$, $\partial\log f(A|\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}$. Specifically, consider the working models $m(A,\boldsymbol{X};\boldsymbol{\alpha}) = \alpha_0 + \alpha_1 A + \boldsymbol{\alpha}_2^T\boldsymbol{X}$ and $f(A|\boldsymbol{X};\boldsymbol{\psi})$ being the normal density with mean $\psi_0 + \boldsymbol{\psi}_1^T\boldsymbol{X}$ and variance ψ_2 ($\psi_2 > 0$). For this particular choice, (4.22) and (4.23) become

$$\mathbf{0} = \sum_{i=1}^n\int_{\mathcal{A}}\{a - E_f(A)\}\left[1 - \frac{I(A_i = a)}{f(a|\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})}\right]\begin{bmatrix}1\\ a\\ \boldsymbol{X}_i\end{bmatrix}f(a)\mathrm{d}a,$$

$$\mathbf{0} = \sum_{i=1}^{n} \{A_i - E_f(A)\} \frac{f(A_i)}{f(A_i|\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})} \begin{bmatrix} \varepsilon_{A,i}(\hat{\psi}_{n,0}^{\mathrm{BR}}, \hat{\boldsymbol{\psi}}_{n,1}^{\mathrm{BR}}) \begin{bmatrix} 1 \\ \boldsymbol{X}_i \end{bmatrix} \\ \varepsilon_{A,i}(\hat{\psi}_{n,0}^{\mathrm{BR}}, \hat{\boldsymbol{\psi}}_{n,1}^{\mathrm{BR}})^2 - \hat{\psi}_{n,2}^{\mathrm{BR}} \end{bmatrix} \times \{Y_i - m(A_i, \boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})\},$$

with $\varepsilon_{A,i}(\psi_0, \boldsymbol{\psi}_1) = A_i - \psi_0 - \boldsymbol{\psi}_1^T \boldsymbol{X}_i$. The estimator $\hat{\beta}_{n,1,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})$, with $\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}$ and $\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}}$ the solutions to (4.22) and (4.23), then defines the bias-reduced doubly robust estimator of β_1. By Theorem 4.1, these nuisance parameter estimators ensure that the doubly robust estimator of β_1 has minimal squared first-order asymptotic bias, although they may not deliver a doubly robust estimator of β_0 with this property. If one is additionally interested in the estimation of β_0, a similar strategy can be used to obtain a bias-reduced doubly robust estimator of β_0, possibly resulting in different estimators for the nuisance parameters.

4.5.2 G-estimation for semiparametric regression models

G-estimation in semiparametric linear regression models

Consider the semiparametric linear regression model

$$E(Y|A, \boldsymbol{X}) = m_0(\boldsymbol{X}) + \tau_0 A$$

(Robins et al. 1992). A doubly robust G-estimator $\hat{\tau}_{n,\mathrm{DR}}^{\mathrm{G}}(\boldsymbol{\psi}, \boldsymbol{\alpha})$ of τ_0 is obtained by solving

$$0 = \sum_{i=1}^{n} U_{\mathrm{G}}(\boldsymbol{O}_i; \tau, \boldsymbol{\psi}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \{Y_i - \tau A_i - m(\boldsymbol{X}_i; \boldsymbol{\alpha})\}\{A_i - \pi(\boldsymbol{X}_i; \boldsymbol{\psi})\},$$

which is unbiased if either the working model $\pi(\boldsymbol{X}; \psi)$ for $E(A|\boldsymbol{X})$, denoted $\mathcal{M}(\boldsymbol{\psi})$, or the working model $m(\boldsymbol{X}; \boldsymbol{\alpha})$ for $E(Y|A = 0, \boldsymbol{X})$, denoted $\mathcal{M}(\boldsymbol{\alpha})$, is correctly specified. For fixed $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$, the doubly robust estimator $\hat{\tau}_{n,\mathrm{DR}}^{\mathrm{G}}(\boldsymbol{\psi}, \boldsymbol{\alpha})$ admits the expansion $n^{1/2}\{\hat{\tau}_{n,\mathrm{DR}}^{\mathrm{G}}(\boldsymbol{\psi}, \boldsymbol{\alpha}) - \tau_0\} = n^{-1/2} \sum_{i=1}^{n} \phi_{\mathrm{G}}(\boldsymbol{O}_i; \tau_0, \boldsymbol{\psi}, \boldsymbol{\alpha})\} + o_p(1)$ with

influence function

$$\phi_{\mathrm{G}}(\boldsymbol{O};\tau_0,\boldsymbol{\psi},\boldsymbol{\alpha}) = \frac{\{A-\pi(\boldsymbol{X};\boldsymbol{\psi})\}\{Y-m(\boldsymbol{X};\boldsymbol{\alpha})\}}{E[A\{A-\pi(\boldsymbol{X};\boldsymbol{\psi})\}]} - \tau_0.$$

The gradients of $\phi_{\mathrm{G}}(\boldsymbol{O};\tau_0,\boldsymbol{\psi},\boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ then define the estimating equations for $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$; that is, we solve $(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})$ from the system

$$\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{n} \{A_i - \pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\} m_{\boldsymbol{\alpha}}(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}}),\\
\mathbf{0} &= \sum_{i=1}^{n} \{Y_i - m(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})\} \widehat{\boldsymbol{W}}_n(A_i,\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}),
\end{aligned}$$

with

$$\begin{aligned}
\widehat{\boldsymbol{W}}_n(A,\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}) =& \{A-\pi(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\} \times n^{-1} \sum_{j=1}^{n} A_j \pi_{\boldsymbol{\psi}}(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\\
&- \pi_{\boldsymbol{\psi}}(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}) \times n^{-1} \sum_{j=1}^{n} A_j \{A_j - \pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\},
\end{aligned}$$

where $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}) = \partial\pi(\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}$, $m_{\boldsymbol{\alpha}}(\boldsymbol{X};\boldsymbol{\alpha}) = \partial m(\boldsymbol{X};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. For a linear outcome model $m(\boldsymbol{X};\boldsymbol{\alpha}) = \alpha_0 + \boldsymbol{\alpha}_1^T\boldsymbol{X}$ and a logistic regression model $\pi(\boldsymbol{X};\boldsymbol{\psi}) = \mathrm{expit}(\psi_0 + \boldsymbol{\psi}_1^T\boldsymbol{X})$ for the propensity score, the estimating equation for $\boldsymbol{\psi}$ reduces to standard MLE because $m_{\boldsymbol{\alpha}}(\boldsymbol{X};\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}}) = (1,\boldsymbol{X}^T)^T$. The estimating equation for $\boldsymbol{\alpha}$ is obtained by substituting $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}) = \pi(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\{1-\pi(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\}(1,\boldsymbol{X}^T)^T$.

G-estimation in semiparametric log-linear regression models

Consider now the semiparametric log-linear model

$$\log E(Y|A,\boldsymbol{X}) = m_0(\boldsymbol{X}) + \tau_0 A$$

(Robins et al. 1992). A doubly robust G-estimator $\hat{\tau}'^{,\mathrm{G}}_{n,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\alpha})$ of τ_0 is obtained by solving the estimating equation

$$0 = \sum_{i=1}^{n} U'_{\mathrm{G}}(\boldsymbol{O}_i;\tau,\boldsymbol{\psi},\boldsymbol{\alpha})\}$$

$$= \sum_{i=1}^{n} \{A_i - \pi(\boldsymbol{X}_i; \boldsymbol{\psi})\}[Y_i \exp(-\tau A_i) - \exp\{m(\boldsymbol{X}_i; \boldsymbol{\alpha})\}],$$

which is unbiased if either the working model ($\mathcal{M}(\boldsymbol{\psi})$) $\pi(\boldsymbol{X}; \boldsymbol{\psi})$ for $E(A|\boldsymbol{X})$ or the working model ($\mathcal{M}(\boldsymbol{\alpha})$) $m(\boldsymbol{X}; \boldsymbol{\alpha})$ for $\log E(Y|A=0, \boldsymbol{X})$ is correctly specified. Although the gradients of the corresponding influence function $\phi'_{\mathrm{G}}(\boldsymbol{O}_i; \tau_0, \boldsymbol{\psi}, \boldsymbol{\alpha}) = -E^{-1}\{\partial U'_{\mathrm{G}}(\boldsymbol{O}; \tau_0, \boldsymbol{\psi}, \boldsymbol{\alpha})/\partial\tau\} U'_{\mathrm{G}}(\boldsymbol{O}_i; \tau_0, \boldsymbol{\psi}, \boldsymbol{\alpha})$ w.r.t. the nuisance parameters continue to deliver consistent estimators for the nuisance parameters, these estimators no longer ensure minimal squared first-order asymptotic bias under misspecification of both working models because these gradients now depend on the unknown population value τ_0. Nevertheless, Theorem 4.1 continues to apply for score tests of the null hypothesis that $\tau = \tilde{\tau}$ for some $\tilde{\tau}$. In particular, when the estimators $\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}$ and $\hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}}$ are defined with the known value $\tilde{\tau}$ substituted for the unknown value of τ, they minimize $E^2\{\phi'_{\mathrm{G}}(\boldsymbol{O}; \tilde{\tau}, \hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\alpha}}_n^{\mathrm{BR}})\}$.

4.5.3 Mean outcome when missingness is non-ignorable

We reconsider the estimation of a population mean outcome $\mu_0 = E(Y)$ in the presence of incomplete data. Suppose, as in Section 3.3, that we have i.i.d. data $\{\boldsymbol{O}_i = (R_i Y_i, R_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$, but that in contrast to Section 3.3, missingness is not ignorable, so that the missing data is missing not at random (NMAR). Suppose therefore that (i) $P(R=1|\boldsymbol{X}, Y) > 0$ with probability one and (ii) if $P(R=0|\boldsymbol{X}) > 0$,

$$P(R=0|\boldsymbol{X}, Y) = \mathrm{expit}\{h_0(\boldsymbol{X}) + q(\boldsymbol{X}, Y; \boldsymbol{\kappa})\}$$

for some unknown function $h_0(\boldsymbol{X})$ and a user-specified selection bias function $q(\boldsymbol{X}, Y; \boldsymbol{\kappa})$ with known $\boldsymbol{\kappa}$ and $q(\boldsymbol{X}, 0; \boldsymbol{\kappa}) \equiv q(\boldsymbol{X}, Y; \mathbf{0}) \equiv 0$, e.g., $q(\boldsymbol{X}, Y; \kappa) = \kappa Y$. Let $\mathcal{M}(\boldsymbol{\kappa})$ denote the model for the full data defined by the assumptions (i) and (ii). Since $\boldsymbol{\kappa} = \mathbf{0}$ encodes MAR, the selection bias function $q(\boldsymbol{X}, Y; \boldsymbol{\kappa})$ encodes the degree of deviation from the MAR assumption (Scharfstein et al. 1999a,b). Scharfstein et al. (1999a) show that for each choice of $\boldsymbol{\kappa}$, model $\mathcal{M}(\boldsymbol{\kappa})$ places no restrictions on the observed data law so that in particular the observed data carry no information of $\boldsymbol{\kappa}$. Model $\mathcal{M}(\boldsymbol{\kappa})$ is hence particularly useful for a sensitivity

analysis based upon varying $\boldsymbol{\kappa}$. It can be shown that at $\mathcal{M}(\boldsymbol{\kappa})$,

$$\mu_0 = E\left[RY/\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa}) + \{1 - R/\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})\}\, m_0(\boldsymbol{X})\right]$$

with

$$\begin{aligned} m_0(\boldsymbol{X}) &= E(Y|R=0,\boldsymbol{X}) \\ &= E[Y\exp\{q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}|R=1,\boldsymbol{X}]/E[\exp\{q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}|R=1,\boldsymbol{X}] \\ \pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa}) &= P(R=1|Y,\boldsymbol{X}) \\ &= [1+\exp\{h_0(\boldsymbol{X}) + q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}]^{-1}. \end{aligned}$$

4

Doubly robust estimation

To construct a doubly robust estimator for the target parameter μ_0 (Scharfstein et al. 1999b), we need two working models: (i) a model $\mathcal{M}(\boldsymbol{\xi})$ for the unknown function $m_0(\boldsymbol{X})$ given by $m(\boldsymbol{X};\boldsymbol{\xi})$, where $m(\boldsymbol{X};\boldsymbol{\xi})$ is a known function smooth in a finite-dimensional parameter $\boldsymbol{\xi}$ and (ii) a model $\mathcal{M}(\boldsymbol{\psi})$ for the unknown function $h_0(\boldsymbol{X})$ given by $h(\boldsymbol{X};\boldsymbol{\psi})$, where $h(\boldsymbol{X};\boldsymbol{\psi})$ is a known function smooth in a finite-dimensional parameter $\boldsymbol{\psi}$. The induced working model for $\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})$ is then $\pi(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa}) = [1+\exp\{h(\boldsymbol{X};\boldsymbol{\psi}) + q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}]^{-1}$. For given $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, the target parameter μ_0 can be estimated as $\hat{\mu}_{n,\mathrm{DR}} \equiv \hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\xi};\boldsymbol{\kappa})$, obtained as a solution to the estimating equation $\sum_{i=1}^n \phi_q(\boldsymbol{O}_i;\hat{\mu}_{n,\mathrm{DR}},\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa})\} = 0$ where

$$\begin{aligned} &\phi_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa}) \\ &\quad = \frac{RY}{\pi(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa})} + \left\{1 - \frac{R}{\pi(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa})}\right\} m(\boldsymbol{X};\boldsymbol{\xi}) - \mu \end{aligned} \tag{4.24}$$

equals the influence function of μ_0 (at fixed nuisance parameters). It is not difficult to see that the estimator $\hat{\mu}_{n,\mathrm{DR}}$ is doubly robust in the sense that it is consistent under model $\mathcal{M}(\boldsymbol{\kappa}) \cap \{\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\xi})\}$. However, using standard methods, it is no longer straightforward how one can estimate $\boldsymbol{\psi}$, because of the dependence of missingness on the incomplete outcome. An alternative class of estimators of $\boldsymbol{\psi}$ has been proposed in Rotnitzky and Robins (1997).

Bias-reduced doubly robust estimation

The bias-reduced doubly robust estimator can be straightforwardly obtained as a consequence of the double robustness of $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi},\boldsymbol{\xi};\boldsymbol{\kappa})$ under model $\mathcal{M}(\boldsymbol{\kappa})\cap\{\mathcal{M}(\boldsymbol{\psi})\cup\mathcal{M}(\boldsymbol{\xi})\}$. Upon taking the gradients of ϕ_q with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$, this amounts to solving $(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ from the system

$$\mathbf{0}=\sum_{i=1}^{n}(1-R_i[1+\exp\{h(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})+q(\boldsymbol{X}_i,Y_i;\boldsymbol{\kappa})\}])m_{\boldsymbol{\xi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\}, \tag{4.25}$$

$$\mathbf{0}=\sum_{i=1}^{n}R_i\{Y_i-m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\}\exp\{h(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})+q(\boldsymbol{X}_i,Y_i;\boldsymbol{\kappa})\}h_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}), \tag{4.26}$$

with $h_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi})=\partial h(\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ and $m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi})=\partial m(\boldsymbol{X};\boldsymbol{\xi})/\partial\boldsymbol{\xi}$. Note that the tilt function $\exp\{q(\boldsymbol{X},Y,\boldsymbol{\kappa})\}$ in the latter involves the selection bias function which links the outcome distribution in the responders to that in the non-responders, thereby assuring the unbiasedness of this estimating equation under model $\mathcal{M}(\boldsymbol{\kappa})\cap\mathcal{M}(\boldsymbol{\xi})$.

As in Section 4.3, whenever $m(\boldsymbol{X};\boldsymbol{\xi})$ lies within the span of the gradient $m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi})$, e.g., for linear models, the restrictions imposed by the estimating equation (4.25) reduce the doubly robust estimator to a simple IPTW estimator $n^{-1}\sum_{i=1}^{n}R_iY_i/\pi(\boldsymbol{X}_i,Y_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}},\boldsymbol{\kappa})$. Additionally, whenever the function $h(\boldsymbol{X};\boldsymbol{\psi})$ includes a constant term, the restrictions implied by (4.26) ensure that the doubly robust estimator reduces to a mean imputation estimator

$$n^{-1}\sum_{i=1}^{n}\left\{R_iY_i+(1-R_i)m(\boldsymbol{X}_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\right\}.$$

Projection estimator of Cao et al. (2009) with non-ignorable missingness

We observed that the results of Section 4.3 immediately extend to non-ignorable missingness, which is in contrast to certain other alternative nuisance parameter estimation strategies. For instance, below we show that when missingness is non-ignorable, the strategy of Cao et al. (2009) does not necessarily lead to an unbiased estimating function for the parameter $\boldsymbol{\xi}$ indexing the working model $m(\boldsymbol{X};\boldsymbol{\xi})$ for $E(Y|R=0,\boldsymbol{X})$.

Suppose $\mathcal{M}(\boldsymbol{\kappa}) \cap \mathcal{M}(\boldsymbol{\psi})$ holds and for simplicity assume that $h(\boldsymbol{X})$ is fully specified and thus equals $h_0(\boldsymbol{X})$. We do not assume $\mathcal{M}(\boldsymbol{\xi})$ necessarily holds. Under these assumptions, the influence function of the doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\xi};\boldsymbol{\kappa})$ is given by $\phi_q(\boldsymbol{O};\mu,\boldsymbol{\xi},\boldsymbol{\kappa})$, given in (4.24), but with the difference that the dependence on $\boldsymbol{\psi}$ is dropped. Following the argument of Cao et al. (2009), the aim is to estimate $\boldsymbol{\xi}$ by minimizing $\mathrm{var}\{\phi_q(\boldsymbol{O};\mu,\boldsymbol{\xi},\boldsymbol{\kappa})\}$. Using the law of iterated variance,

$$\begin{aligned}&\mathrm{var}\{\phi_q(\boldsymbol{O};\mu,\boldsymbol{\xi},\boldsymbol{\kappa})\}\\&\quad = E[\mathrm{var}\{\phi_q(\boldsymbol{O};\mu,\boldsymbol{\xi},\boldsymbol{\kappa})|\boldsymbol{X},Y\}] + \mathrm{var}[E\{\phi_q(\boldsymbol{O};\mu,\boldsymbol{\xi},\boldsymbol{\kappa})|\boldsymbol{X},Y\}],\end{aligned}$$

this variance equals

$$E\left[\frac{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}{\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}\{Y-m(\boldsymbol{X};\boldsymbol{\xi})\}^2\right] + \mathrm{var}(Y), \tag{4.27}$$

where $\{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})\}/\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa}) = \exp\{h_0(\boldsymbol{X})+q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}$. The aim is then to find the value $\boldsymbol{\xi}^*$ and a corresponding estimator $\hat{\boldsymbol{\xi}}_n$ with $\mathrm{plim}(\hat{\boldsymbol{\xi}}_n) = \boldsymbol{\xi}^*$ such that $\boldsymbol{\xi}^*$ minimizes (4.27). Taking the gradient of (4.27) with respect to $\boldsymbol{\xi}$ yields that $\boldsymbol{\xi}^*$ should solve

$$E\left[\frac{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}{\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}\{Y-m(\boldsymbol{X};\boldsymbol{\xi}^*)\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}^*)\right] = \mathbf{0}.$$

However, when $m(\boldsymbol{X};\boldsymbol{\xi})$ is correctly specified, $\boldsymbol{\xi}^*$ should equal $\boldsymbol{\xi}_0$ where $\boldsymbol{\xi}_0$ is such that $m_0(\boldsymbol{X}) = m(\boldsymbol{X};\boldsymbol{\xi}_0)$. We show below that this is not generally the case. The left-hand side of the latter equation evaluated at $\boldsymbol{\xi}_0$ can be written as

$$\begin{aligned}&E\left[R\frac{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}{\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}\{Y-m(\boldsymbol{X};\boldsymbol{\xi}_0)\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)\right]\\&\qquad + E\left[(1-R)\frac{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}{\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})}\{Y-m(\boldsymbol{X};\boldsymbol{\xi}_0)\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)\right]\\&= E\left[\{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\kappa})\}\{Y-m(\boldsymbol{X};\boldsymbol{\xi}_0)\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)\right]\\&\qquad + E\left[(1-R)\exp\{h_0(\boldsymbol{X})+q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}\{Y-m(\boldsymbol{X};\boldsymbol{\xi}_0)\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)\right]\end{aligned}$$

$$= E[(1-R)\{Y - m_0(\boldsymbol{X})\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)] \tag{4.28}$$

$$+ E\Big[(1-R)E[\exp\{q(\boldsymbol{X},Y,\boldsymbol{\kappa})\}\{Y - m_0(\boldsymbol{X})\}|R=0,\boldsymbol{X}] \tag{4.29}$$

$$\times \exp\{h_0(\boldsymbol{X})\}m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)\Big]. \tag{4.30}$$

Expression (4.28) equals $\mathbf{0}$ because $m_0(\boldsymbol{X}) = E(Y|R=0,\boldsymbol{X})$. Expression (4.29)-(4.30) can be written as

$$E\{(1-R)\text{cov}[\exp\{q(\boldsymbol{X},Y;\boldsymbol{\kappa}\},Y|R=0,\boldsymbol{X}]h_0(\boldsymbol{X})m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}_0)\},$$

which does not generally equal $\mathbf{0}$. This shows that in general, $\boldsymbol{\xi}^*$ will not equal $\boldsymbol{\xi}_0$, even when $\mathcal{M}(\boldsymbol{\xi})$ holds.

Now suppose that $h_0(\boldsymbol{X})$ is unknown but known to follow a parametric model $h(\boldsymbol{X};\boldsymbol{\psi})$. An estimator $\hat{\boldsymbol{\psi}}_n$ for $\boldsymbol{\psi}_0$ cannot be obtained using standard methods. An estimator $\hat{\boldsymbol{\psi}}_n$ can however be obtained by solving the estimating equations $\sum_{i=1}^n \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\hat{\boldsymbol{\psi}}_n,\boldsymbol{\rho},\boldsymbol{\kappa})\} = \mathbf{0}$ with

$$\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\rho},\boldsymbol{\kappa}) = (1 - R[1 + \exp\{h(\boldsymbol{X};\boldsymbol{\psi}) + q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}])\boldsymbol{\rho}(\boldsymbol{X})$$

where $\boldsymbol{\rho}$ is an arbitrary function of $\boldsymbol{X}$ of the same dimension as $\boldsymbol{\psi}$. One could for instance use the estimator $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ where

$$\partial\phi_q(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa})/\partial\boldsymbol{\xi} = \boldsymbol{U}_{\boldsymbol{\psi}}\{\boldsymbol{O};\boldsymbol{\psi},m_{\boldsymbol{\xi}}(\boldsymbol{X};\boldsymbol{\xi}),\boldsymbol{\kappa}\}.$$

At $\mathcal{M}(\boldsymbol{\kappa})\cap\mathcal{M}(\boldsymbol{\psi})$, the influence function of $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n,\boldsymbol{\xi};\boldsymbol{\kappa})$ is given by

$$\tilde{\phi}_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa}) = \phi_q(\boldsymbol{Z};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa}) - \boldsymbol{c}^{*,T}(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\rho},\boldsymbol{\kappa})$$

with

$$\boldsymbol{c}^{*,T}(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa}) = E\left\{\partial\phi_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa})/\partial\boldsymbol{\psi}^T\right\} \times E^{-1}\left\{\partial\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\rho},\boldsymbol{\kappa})/\partial\boldsymbol{\psi}^T\right\},$$

$$E\left\{\partial\phi_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa})/\partial\boldsymbol{\psi}^T\right\} = E\Big[\{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa})\} \times \{m_0(\boldsymbol{X}) - m(\boldsymbol{X};\boldsymbol{\xi})\}h_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi})\Big]$$

$$E\left\{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\rho},\boldsymbol{\kappa})/\partial \boldsymbol{\psi}^T\right\} = E\Big[\{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa})\}\boldsymbol{\rho}(\boldsymbol{X})h^T_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi})\Big].$$

The influence function can be rewritten as

$$\begin{aligned}
&\tilde{\phi}_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\kappa)\\
&\quad = RY[1+\exp\{h(\boldsymbol{X};\boldsymbol{\psi})+q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}]\\
&\qquad + (1-R[1+\exp\{h(\boldsymbol{X};\boldsymbol{\psi})+q(\boldsymbol{X},Y;\boldsymbol{\kappa})\}])\widetilde{m}\left\{\widetilde{\boldsymbol{X}};\boldsymbol{\xi},\boldsymbol{c}^*(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})\right\}
\end{aligned}$$

with $\widetilde{\boldsymbol{X}} = \{\boldsymbol{X}^T,\boldsymbol{\rho}^T(\boldsymbol{X})\}^T$ and the extended model

$$\widetilde{m}\left\{\widetilde{\boldsymbol{X}};\boldsymbol{\xi},\boldsymbol{c}^*(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})\right\} = m(\boldsymbol{X};\boldsymbol{\xi}) - \boldsymbol{c}^{*,T}(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})\boldsymbol{\rho}(\boldsymbol{X}).$$

The variance of $\tilde{\phi}_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})$ equals

$$E\left(\frac{1-\pi_0(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa})}{\pi_0(\boldsymbol{X},Y;\boldsymbol{\psi},\boldsymbol{\kappa})}\left[Y-\widetilde{m}\left\{\widetilde{\boldsymbol{X}};\boldsymbol{\xi},\boldsymbol{c}^*(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})\right\}\right]^2\right)+\text{var}(Y). \qquad (4.31)$$

In principle, one could then jointly solve $(\hat{\boldsymbol{\xi}}_n^T,\hat{\boldsymbol{c}}_n^T)^T$ from estimating equations based on the gradient with respect to $(\boldsymbol{\xi}^T,\boldsymbol{c}^T)^T$ of (4.31) with $\boldsymbol{c}^*(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})$ replaced by $\boldsymbol{c}$. However, plim$(\hat{\boldsymbol{c}}_n)$ would not equal the value $\boldsymbol{c}^*(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})$ because $\boldsymbol{c}^{*T}(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\rho},\boldsymbol{\kappa})$ does not correspond to the orthogonal projection of the function $\phi_q(\boldsymbol{O};\mu,\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\kappa})$ onto the linear space spanned by $\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\rho},\boldsymbol{\kappa})$ so that $\boldsymbol{c}^*(\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{\kappa})$ cannot be estimated via (weighted) least squares regression. We may conclude that, even when making the unrealistic assumption that cov$[\exp\{q(\boldsymbol{X},Y;\boldsymbol{\kappa}\},Y|R=0,\boldsymbol{X}]=0$ holds with probability one so that $\boldsymbol{\xi}^*=\boldsymbol{\xi}_0$, it is not straightforward to find an estimator $\hat{\boldsymbol{\xi}}_n$ with probability limit $\boldsymbol{\xi}^*$ that minimizes (4.31) and thus prohibiting the projection construction as in Cao et al. (2009).

4.5.4 Multiply robust estimation in semiparametric interaction models

The principle behind the biased-reduced doubly robust estimation strategy is extensible to certain multiply robust estimators, estimators that are consistent under a union model that assumes that at least one of several working models holds. Consider i.i.d. data $\{\boldsymbol{O}_i = (Y_i, \boldsymbol{A}_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$, where Y_i is the outcome, $\boldsymbol{A}_i = (A_{i1}, A_{i2})^T$ is a vector of binary exposure variables and $\boldsymbol{X}_i$ is a vector of extraneous variables. Vansteelandt et al. (2008) develop inference for β_0 in the semiparametric interaction model $\mathcal{M}$ defined by

$$E(Y|\boldsymbol{A}, \boldsymbol{X}) = \beta_0 A_1 A_2 + q_1(A_1, \boldsymbol{X}) + q_2(A_2, \boldsymbol{X}) + q_0(\boldsymbol{X}), \tag{4.32}$$

where $q_j(A_j, \boldsymbol{X})$ $(j = 1, 2)$ and $q_0(\boldsymbol{X})$ are unknown functions satisfying $q_j(0, \boldsymbol{X}) = 0$. Vansteelandt et al. (2008) show how a multiply robust estimator for β_0 can be obtained under this model.

Conditionally independent exposures

Suppose for now that the exposures A_1 and A_2 are conditionally independent given $\boldsymbol{X}$. Let $\mathcal{M}_{\text{cip}}$ be the model defined by the model $\mathcal{M}$ and the assumption $A_1 \perp\!\!\!\perp A_2 | \boldsymbol{X}$. Let $\mathcal{M}(\boldsymbol{\psi}^{(j)})$ be a working model for $E(A_j|\boldsymbol{X})$, e.g., $\pi_j(\boldsymbol{X}; \boldsymbol{\psi}^{(j)}) = \text{expit}(\psi_1^{(j)} + \boldsymbol{\psi}_2^{(j),T} \boldsymbol{X})$ $(j = 1, 2)$. Furthermore, let $\mathcal{M}(\boldsymbol{\alpha}^{(j)})$ be a working model for the main effect $q_j(A_j, \boldsymbol{X})$, e.g., $q_j(A_j, \boldsymbol{X}; \boldsymbol{\alpha}^{(j)}) = A_j(\alpha_1^{(j)} + \boldsymbol{\alpha}_2^{(j),T} \boldsymbol{X})$ $(j = 1, 2)$ and finally, let $\mathcal{M}(\boldsymbol{\alpha}^{(0)})$ be a working model for the covariate effect $q_0(\boldsymbol{X})$, e.g., $q_0(\boldsymbol{X}; \boldsymbol{\alpha}^{(0)}) = \alpha_1^{(0)} + \boldsymbol{\alpha}_2^{(0),T} \boldsymbol{X}$.

A multiply (specifically, quadruply) robust estimator $\hat{\beta}_{n,\text{cip}} \equiv \hat{\beta}_{n,\text{cip}}(\boldsymbol{\psi}, \boldsymbol{\alpha})$ for β_0 under the union model $\mathcal{M}_{\text{cip}} \cap [\{\mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\alpha}^{(0)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$ can be obtained as the solution to an estimating equation of the form (Vansteelandt et al. 2008)

$$0 = \sum_{i=1}^{n} U\left(\boldsymbol{O}_i; \hat{\beta}_{n,\text{cip}}, \boldsymbol{\psi}, \boldsymbol{\alpha}\right)$$

$$= \sum_{i=1}^{n} \Big[\{A_{i1} - \pi_1(\boldsymbol{X}_i; \boldsymbol{\psi}^{(1)})\}\{A_{i2} - \pi_2(\boldsymbol{X}_i; \boldsymbol{\psi}^{(2)})\} \\ \times \big\{ Y_i - \hat{\beta}_{n,\text{cip}} A_{i1} A_{i2} - q_1(A_{i1}, \boldsymbol{X}_i; \boldsymbol{\alpha}^{(1)}) - q_2(A_{i2}, \boldsymbol{X}_i; \boldsymbol{\alpha}^{(2)}) - q_0(\boldsymbol{X}_i; \boldsymbol{\alpha}^{(0)}) \big\} \Big],$$

with $\boldsymbol{\psi} = (\boldsymbol{\psi}^{(1),T}, \boldsymbol{\psi}^{(2),T})^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(0),T}, \boldsymbol{\alpha}^{(1),T}, \boldsymbol{\alpha}^{(2),T})^T$. It is easy to see that for fixed values of the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$, the quadruply robust estimator $\hat{\beta}_{n,\text{cip}}$ admits the expansion $n^{1/2}(\hat{\beta}_{n,\text{cip}} - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \phi_{\text{cip}}(\boldsymbol{O}_i; \beta_0, \boldsymbol{\psi}, \boldsymbol{\alpha}) + o_p(1)$, with influence function

$$\phi_{\text{cip}}(\boldsymbol{O}; \beta_0, \boldsymbol{\psi}, \boldsymbol{\alpha}) \\ = \frac{\{A_1 - \pi_1(\boldsymbol{X}; \boldsymbol{\psi}^{(1)})\}\{A_2 - \pi_2(\boldsymbol{X}; \boldsymbol{\psi}^{(2)})\}}{E[A_1 A_2 \{A_1 - \pi_1(\boldsymbol{X}; \boldsymbol{\psi}^{(1)})\}\{A_2 - \pi_2(\boldsymbol{X}; \boldsymbol{\psi}^{(2)})\}]} \\ \times \big\{ Y - q_1(A_1, \boldsymbol{X}; \boldsymbol{\alpha}^{(1)}) - q_2(A_2, \boldsymbol{X}; \boldsymbol{\alpha}^{(2)}) - q_0(\boldsymbol{X}; \boldsymbol{\alpha}^{(0)}) \big\} - \beta_0.$$

In what follows, with a slight abuse of notation, we will also denote $\phi_{\text{cip}}(\boldsymbol{O}; \beta, \boldsymbol{\psi}, \boldsymbol{\alpha})$ as $\phi_{\text{cip}}(\boldsymbol{O}; \beta, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \boldsymbol{\alpha})$ or $\phi_{\text{cip}}(\boldsymbol{O}; \beta, \boldsymbol{\psi}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)})$. Furthermore, assume that the union model $\mathcal{M}_{\text{cip}} \cap [\{\mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\alpha}^{(0)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$ is correctly specified.

Suppose first that the working model $\mathcal{M}(\boldsymbol{\alpha}^{(1)})$ is misspecified, in which case the union model $\mathcal{M}_{\text{cip}} \cap [\{\mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$ must hold. Then $\phi_{\text{cip}}(\boldsymbol{O}; \beta_0, \boldsymbol{\psi}^*, \boldsymbol{\alpha}^{(0),*}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2),*})$ has mean zero for all $\boldsymbol{\alpha}^{(1)}$ so that its gradient w.r.t. $\boldsymbol{\alpha}^{(1)}$ also has mean zero under model $\mathcal{M}_{\text{cip}} \cap [\{\mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$. Because under this union model, $\mathcal{M}(\boldsymbol{\psi}^{(2)})$ is always correctly specified, this leads to an unbiased estimating function for $\boldsymbol{\psi}^{(2)}$. Specifically, this leads to the estimating equation

$$\mathbf{0} = \sum_{i=1}^{n} A_{i1} \{A_{i1} - \pi_1(\boldsymbol{X}_i; \boldsymbol{\psi}^{(1)})\}\{A_{i2} - \pi_2(\boldsymbol{X}_i; \boldsymbol{\psi}^{(2)})\} \begin{bmatrix} 1 \\ \boldsymbol{X}_i \end{bmatrix},$$

which is unbiased under $\mathcal{M}(\boldsymbol{\psi}^{(2)})$. The solution $\hat{\boldsymbol{\psi}}^{(2)}_{n,\text{BR}}$ is thus a consistent estimator for $\boldsymbol{\psi}^{(2)}$. An estimator $\hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR}}$ for $\boldsymbol{\psi}^{(1)}$ is likewise constructed by taking the gradient of $\phi_{\text{cip}}(\boldsymbol{O}; \beta_0, \boldsymbol{\psi}^*, \boldsymbol{\alpha}^{(0),*}, \boldsymbol{\alpha}^{(1),*}, \boldsymbol{\alpha}^{(2)})$ w.r.t. $\boldsymbol{\alpha}^{(2)}$.

Suppose next that the model $\mathcal{M}(\boldsymbol{\psi}^{(1)})$ is misspecified, in which case the union model $\mathcal{M}_{\text{cip}} \cap [\{\mathcal{M}(\boldsymbol{\alpha}^{(0)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$ must hold. Then $\phi_{\text{cip}}(\boldsymbol{O}; \beta_0, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2),*}, \boldsymbol{\alpha}^*)$ has mean zero for all $\boldsymbol{\psi}^{(1)}$, so that its gradient w.r.t. $\boldsymbol{\psi}^{(1)}$ has mean zero under model $\mathcal{M}_{\text{cip}} \cap [\{\mathcal{M}(\boldsymbol{\alpha}^{(0)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$. Because under this union model, $\mathcal{M}(\boldsymbol{\alpha}^{(2)})$ is always correctly specified, this leads to an unbiased estimating function for $\boldsymbol{\alpha}^{(2)}$. Specifically, this leads to the estimating equation

$$\mathbf{0} = \sum_{i=1}^{n} \Big[\{Y_i - q_1(A_{i1}, \boldsymbol{X}_i; \boldsymbol{\alpha}^{(1)}) - q_2(A_{i2}, \boldsymbol{X}_i; \boldsymbol{\alpha}^{(2)}) - q_0(\boldsymbol{X}_i; \boldsymbol{\alpha}^{(0)})\} \times \widehat{\boldsymbol{W}}_n^{(2)}(A_{i1}, A_{i2}, \boldsymbol{X}_i; \boldsymbol{\psi}) \Big]$$

with

$$\begin{aligned}
&\widehat{\boldsymbol{W}}_n^{(2)}(A_1, A_2, \boldsymbol{X}; \boldsymbol{\psi}) \\
&= \{A_2 - \pi_2(\boldsymbol{X}; \boldsymbol{\psi}^{(2)})\} \\
&\quad \times \Bigg(\{A_1 - \pi_1(\boldsymbol{X}; \boldsymbol{\psi}^{(1)})\} n^{-1} \sum_{j=1}^{n} \Bigg[A_{j1} A_{j2} \pi_1(\boldsymbol{X}_j; \boldsymbol{\psi}^{(1)}) \{1 - \pi_1(\boldsymbol{X}_j; \boldsymbol{\psi}^{(1)})\} \\
&\qquad \times \{A_{j2} - \pi_2(\boldsymbol{X}_j; \boldsymbol{\psi}^{(2)})\} \begin{bmatrix} 1 \\ \boldsymbol{X}_j \end{bmatrix} \Bigg] \\
&\qquad - \pi_1(\boldsymbol{X}; \boldsymbol{\psi}^{(1)}) \{1 - \pi_1(\boldsymbol{X}; \boldsymbol{\psi}^{(1)})\} \begin{bmatrix} 1 \\ \boldsymbol{X} \end{bmatrix} \\
&\qquad \times \sum_{j=1}^{n} \Big[A_{j1} A_{j2} \{A_{j1} - \pi_1(\boldsymbol{X}_j; \boldsymbol{\psi}^{(1)})\} \{A_{j2} - \pi_2(\boldsymbol{X}_j; \boldsymbol{\psi}^{(2)})\} \Big] \Bigg).
\end{aligned}$$

The solution $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(2)}$ is thus a consistent estimator for $\boldsymbol{\alpha}^{(2)}$. An estimator $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(1)}$ for $\boldsymbol{\alpha}^{(1)}$ is likewise constructed by taking the gradient of $\phi_{\text{cip}}(\boldsymbol{O}; \beta_0, \boldsymbol{\psi}^{(1),*}, \boldsymbol{\psi}^{(2)}, \boldsymbol{\alpha}^*)$ w.r.t. $\boldsymbol{\psi}^{(2)}$.

Because the influence function $\phi_{\text{cip}}(\boldsymbol{O}; \beta_0, \boldsymbol{\psi}, \boldsymbol{\alpha})$ is linear in the target parameter β_0, the result of Theorem 4.1 can be generalized to this multiply robust estimator. Hence, the estimators $\hat{\boldsymbol{\psi}}_{n,\text{BR}}^{(1)}$, $\hat{\boldsymbol{\psi}}_{n,\text{BR}}^{(2)}$, $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(1)}$ and $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(2)}$ locally minimize the squared first-order asymptotic bias of the multiply robust estimator $\hat{\beta}_{n,\text{cip}}(\hat{\boldsymbol{\psi}}_{n,\text{BR}}, \hat{\boldsymbol{\alpha}}_{n,\text{BR}})$.

Conditionally dependent exposures

The above estimation strategy works when $\boldsymbol{\alpha}^{(0)}$ is assumed to be known, e.g., if $\boldsymbol{\alpha}^{(0)}$ is set to $\mathbf{0}$. When $\boldsymbol{\alpha}^{(0)}$ is unknown, progress can be made by relaxing the assumption that $A_1 \perp\!\!\!\perp A_2|\boldsymbol{X}$ via a model, $\mathcal{M}(\boldsymbol{\psi}^{(0)})$, for the conditional log odds ratio function

$$\rho(\boldsymbol{A},\boldsymbol{X}) = \log\left[\frac{f(A_2=1|A_1=1,\boldsymbol{X})/f(A_2|A_1=1,\boldsymbol{X})}{f(A_2=1|A_1,\boldsymbol{X})/f(A_2|A_1,\boldsymbol{X})}\right],$$

4

e.g., $\rho(\boldsymbol{A},\boldsymbol{X};\boldsymbol{\psi}^{(0)}) = (\psi_1^{(0)} + \boldsymbol{\psi}_2^{(0)T}\boldsymbol{X})(1-A_1)(1-A_2)$. This relaxation should not be viewed as a limitation, as it yields a more efficient estimator of β_0 in large samples, even when in truth $A_1 \perp\!\!\!\perp A_2|\boldsymbol{X}$ (Vansteelandt et al. 2008). When making this relaxation, we wish to treat the exposures A_1 and A_2 in a symmetric way; in particular, we wish $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$ to have the same dimension, and likewise $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ to have the same dimension. We therefore follow a different proposal than that used in Vansteelandt et al. (2008).

Redefine the working models $\mathcal{M}(\boldsymbol{\psi}^{(1)})$ and $\mathcal{M}(\boldsymbol{\psi}^{(2)})$ given by $\pi_j(\boldsymbol{X};\boldsymbol{\psi}^{(j)}) = \text{expit}(\psi_1^{(j)} + \boldsymbol{\psi}_2^{(j)T}\boldsymbol{X})$ $(j=1,2)$ to be working models for $E(A_1|A_2=1,\boldsymbol{X})$ and $E(A_2|A_1=1,\boldsymbol{X})$ respectively. Together with $\mathcal{M}(\boldsymbol{\psi}^{(0)})$, these induce compatible models $f(A_1|A_2,\boldsymbol{X};\boldsymbol{\psi}^{(0)},\boldsymbol{\psi}^{(1)})$ and $f(A_2|A_1,\boldsymbol{X};\boldsymbol{\psi}^{(0)},\boldsymbol{\psi}^{(2)})$ for the conditional densities $f(A_1|A_2,\boldsymbol{X})$ and $f(A_2|A_1,\boldsymbol{X})$ (Chen 2007). Next, define arbitrary but fixed conditional density functions $f^*(\boldsymbol{A}|\boldsymbol{X}) = f^*(A_1|\boldsymbol{X})f^*(A_2|\boldsymbol{X})$ with $f^*(A_j|\boldsymbol{X})$ equal to $\pi_j^*(\boldsymbol{X})^{A_j}\{1-\pi_j^*(\boldsymbol{X})\}^{1-A_j}$ $(j=1,2)$ where $\pi_j^*(\boldsymbol{X})$ can be replaced by its maximum likelihood estimate under a parametric model. Now define the estimating function $U^*(\boldsymbol{O};\beta,\boldsymbol{\alpha})$ like $U(\boldsymbol{O};\beta,\boldsymbol{\psi},\boldsymbol{\alpha})$ but with π_j^* instead of π_j. It then follows from Vansteelandt et al. (2008) and Proposition 4.2 below that a multiply robust estimator $\hat{\beta}_{n,\text{ext}} \equiv \hat{\beta}_{n,\text{ext}}(\boldsymbol{\psi},\boldsymbol{\alpha})$ of β_0 under the union model $\mathcal{M} \cap [\{\mathcal{M}(\boldsymbol{\psi}^{(0)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\alpha}^{(0)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(0)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(1)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(1)})\} \cup \{\mathcal{M}(\boldsymbol{\psi}^{(0)}) \cap \mathcal{M}(\boldsymbol{\psi}^{(2)}) \cap \mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$ can be obtained as the solution to

$$0 = \sum_{i=1}^{n} U_{\text{ext}}(\boldsymbol{O}_i;\hat{\beta}_{n,\text{ext}},\boldsymbol{\psi},\boldsymbol{\alpha})\}$$

$$= \sum_{i=1}^{n} \frac{f^*(\boldsymbol{A}_i|\boldsymbol{X}_i)}{f(\boldsymbol{A}_i|\boldsymbol{X}_i;\boldsymbol{\psi})} U^*(\boldsymbol{O}_i;\hat{\beta}_{n,\text{ext}},\boldsymbol{\alpha}),$$

with $\boldsymbol{\psi}$ now denoting $(\boldsymbol{\psi}^{(0),T},\boldsymbol{\psi}^{(1),T},\boldsymbol{\psi}^{(2),T})^T$ where the subscript *ext* indicates that the estimator is defined under the extended model. For fixed values of the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$, the estimator $\hat{\beta}_{n,\text{ext}}$ admits the expansion $n^{1/2}(\hat{\beta}_{n,\text{ext}}-\beta_0) = n^{-1/2}\sum_{i=1}^{n}\phi_{\text{ext}}(\boldsymbol{O}_i;\beta_0,\boldsymbol{\psi},\boldsymbol{\alpha})\} + o_p(1)$ with influence function

$$\begin{aligned}
&\phi_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi},\boldsymbol{\alpha}) \\
&= \frac{\{A_1-\pi_1^*(\boldsymbol{X})\}\{A_2-\pi_2^*(\boldsymbol{X})\}}{E\left[A_1A_2\{A_1-\pi_1^*(\boldsymbol{X})\}\{A_2-\pi_2^*(\boldsymbol{X})\}f^*(\boldsymbol{A}|\boldsymbol{X})/f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi})\right]} \\
&\quad \times \frac{f^*(\boldsymbol{A}|\boldsymbol{X})}{f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi})}\{Y-q_1(A_1,\boldsymbol{X};\boldsymbol{\alpha}^{(1)})-q_2(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(2)})-q_0(\boldsymbol{X};\boldsymbol{\alpha}^{(0)})\}-\beta_0.
\end{aligned}$$

The proof of the proposition below shows the multiply robustness (and more specifically quadruply robustness) property of the estimator $\hat{\beta}_{n,\text{ext}}$.

Proposition 4.2. *The estimator $\hat{\beta}_{n,ext}(\boldsymbol{\psi},\boldsymbol{\alpha})$, defined as the solution to the estimating equation $\sum_{i=1}^{n} U_{ext}(\boldsymbol{O}_i;\hat{\beta}_{n,ext},\boldsymbol{\psi},\boldsymbol{\alpha}) = 0$, is multiply, more specifically quadruply, robust under the model $\mathcal{M}\cap[\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\}\cup\{\mathcal{M}(\boldsymbol{\alpha}^{(0)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$.*

Proof. To prove the quadruply robustness of the estimator $\hat{\beta}_{n,\text{ext}}(\boldsymbol{\psi},\boldsymbol{\alpha})$, we need to show the unbiasedness of the estimating function $U_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*)$ under $\mathcal{M}\cap[\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\}\cup\{\mathcal{M}(\boldsymbol{\alpha}^{(0)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$ with nuisance parameters $\boldsymbol{\psi} = (\boldsymbol{\psi}^{(0),T},\boldsymbol{\psi}^{(1),T},\boldsymbol{\psi}^{(2),T})^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(0),T},\boldsymbol{\alpha}^{(1),T},\boldsymbol{\alpha}^{(2),T})^T$.

Under model $\mathcal{M}\cap\{\mathcal{M}(\boldsymbol{\alpha}^{(0)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}$, $q_0(\boldsymbol{X}) = q_0(\boldsymbol{X};\boldsymbol{\alpha}^{(0),*})$, $q_1(A_1,\boldsymbol{X}) = q_1(A_1,\boldsymbol{X};\boldsymbol{\alpha}^{(1),*})$ and $q_2(A_2,\boldsymbol{X}) = q_0(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(2),*})$. The unbiasedness then trivially follows from the law of iterated expectation.

Under model $\mathcal{M}\cap\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\}$, $f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi}^*)$ equals the true conditional density function of $\boldsymbol{A}$ given $\boldsymbol{X}$. From the law of conditional

expectation, it follows that

$$E\{U_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*)\} = E\left[\frac{f^*(\boldsymbol{A}|\boldsymbol{X})}{f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi}^*)}\{A_1-\pi_1^*(\boldsymbol{X})\}\{A_2-\pi_2^*(\boldsymbol{X})\}\Delta Q_1(\boldsymbol{A},\boldsymbol{X};\boldsymbol{\alpha}^*)\right],$$

with

$$\begin{aligned}\Delta Q_1(\boldsymbol{A},\boldsymbol{X};\boldsymbol{\alpha}^*) &= q_0(\boldsymbol{X})+q_1(A_1,\boldsymbol{X})+q_2(A_2,\boldsymbol{X})-q_0(\boldsymbol{X};\boldsymbol{\alpha}^{(0),*})\\ &\quad -q_1(A_1,\boldsymbol{X};\boldsymbol{\alpha}^{(1),*})-q_2(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(2),*}).\end{aligned}$$

For any function $\omega(\boldsymbol{A},\boldsymbol{X})$ of $\boldsymbol{A}$ and $\boldsymbol{X}$, it holds that

$$\begin{aligned}E\left\{\frac{f^*(\boldsymbol{A}|\boldsymbol{X})}{f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi}^*)}\omega(\boldsymbol{A},\boldsymbol{X})\middle|\boldsymbol{X}\right\} &= \sum_{\boldsymbol{a}}\frac{f^*(\boldsymbol{a}|\boldsymbol{X})}{f(\boldsymbol{a}|\boldsymbol{X};\boldsymbol{\psi}^*)}\omega(\boldsymbol{a},\boldsymbol{X})f(\boldsymbol{a}|\boldsymbol{X};\boldsymbol{\psi}^*)\\ &= E^*\{\omega(\boldsymbol{A},\boldsymbol{X})|\boldsymbol{X}\},\end{aligned}$$

with $E^*(\cdot)$ the expectation taken with respect to $f^*(\boldsymbol{A}|\boldsymbol{X})$. This now implies that $E\{U_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*)\} = E^*\{U^*(\boldsymbol{O};\beta_0,\boldsymbol{\alpha}^*)\}$. The unbiasedness then follows as in Vansteelandt et al. (2008).

Next assume model $\mathcal{M}\cap\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\}$ holds. Because model $\mathcal{M}(\boldsymbol{\alpha}^{(1)})$ holds, we have $q_1(A_1,\boldsymbol{X}) = q_1(A_1,\boldsymbol{X};\boldsymbol{\alpha}^{(1),*})$. From the law of iterated expectation, it then follows that $E\{U_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*)\}$ equals

$$E\left[\frac{f^*(\boldsymbol{A}|\boldsymbol{X})}{f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi}^*)}\{A_1-\pi_1^*(\boldsymbol{X})\}\{A_2-\pi_2^*(\boldsymbol{X})\}\Delta Q_2(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(0),*},\boldsymbol{\alpha}^{(2),*})\right]$$

with

$$\begin{aligned}\Delta Q_2(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(0),*},\boldsymbol{\alpha}^{(2),*}) &= q_2(A_2,\boldsymbol{X})+q_0(\boldsymbol{X})\\ &\quad -q_2(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(2),*})-q_0(\boldsymbol{X};\boldsymbol{\alpha}^{(0),*}).\end{aligned}$$

We can write $f(\boldsymbol{A}|\boldsymbol{X};\boldsymbol{\psi}^*)$ as $f(A_1|A_2,\boldsymbol{X};\boldsymbol{\psi}^{(0),*},\boldsymbol{\psi}^{(1),*})f(A_2|\boldsymbol{X};\boldsymbol{\psi}^*)$. Because both working models $\mathcal{M}(\boldsymbol{\psi}^{(0)})$ and $\mathcal{M}(\boldsymbol{\psi}^{(1)})$ hold, $f(A_1|A_2,\boldsymbol{X};\boldsymbol{\psi}^{(0),*},\boldsymbol{\psi}^{(1),*})$ will equal $f(A_1|A_2,\boldsymbol{X})$. From a similar reasoning as for the previous case, we obtain that

$E\{U_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*)\}$ equals

$$E\left[\frac{f^*(A_2|\boldsymbol{X})\{A_2-\pi_2^*(\boldsymbol{X})\}}{f(A_2|\boldsymbol{X};\boldsymbol{\psi}^*)}\Delta Q_2(A_2,\boldsymbol{X};\boldsymbol{\alpha}^{(0),*},\boldsymbol{\alpha}^{(2),*})E^*\{A_1-\pi_1^*(\boldsymbol{X})|\boldsymbol{X}\}\right]=0$$

because $E^*\{A_1-\pi_1^*(\boldsymbol{X})|\boldsymbol{X}\}=0$.

The unbiasedness of the estimating function $U_{\text{ext}}(\boldsymbol{O};\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*)$ under model $\mathcal{M}\cap\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}$ can be demonstrated analogously. □

4

Estimating functions for $\boldsymbol{\psi}^{(1)}$, $\boldsymbol{\psi}^{(2)}$, $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ are obtained using the same strategy as before but resulting in different estimating functions. To develop an estimating function for $\boldsymbol{\psi}^{(0)}$, suppose $\mathcal{M}(\boldsymbol{\alpha}^{(0)})$ is misspecified but the model $\mathcal{M}\cap[\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}$ holds. In this case, the influence function $\phi_{\text{ext}}(\boldsymbol{O}_i;\beta_0,\boldsymbol{\psi}^*,\boldsymbol{\alpha}^{(0)},\boldsymbol{\alpha}^{(1),*},\boldsymbol{\alpha}^{(2),*})$ has mean zero for all $\boldsymbol{\alpha}^{(0)}$ so that its gradient w.r.t. $\boldsymbol{\alpha}^{(0)}$ has mean zero under $\mathcal{M}\cap[\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(1)})\}\cup\{\mathcal{M}(\boldsymbol{\psi}^{(0)})\cap\mathcal{M}(\boldsymbol{\psi}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\}]$. Because under this union model, $\mathcal{M}(\boldsymbol{\psi}^{(0)})$ is always correctly specified, this leads to an unbiased estimating function for $\boldsymbol{\psi}^{(0)}$ indexing the working model $\mathcal{M}(\boldsymbol{\psi}^{(0)})$. Finally, to develop an estimating function for $\boldsymbol{\alpha}^{(0)}$, suppose that $\mathcal{M}(\boldsymbol{\psi}^{(0)})$ is misspecified but that model $\mathcal{M}\cap\{\mathcal{M}(\boldsymbol{\alpha}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(0)})\}$ holds. Then the influence function $\phi_{\text{ext}}(\boldsymbol{O}_i;\beta_0,\boldsymbol{\psi}^{(0)},\boldsymbol{\psi}^{(1),*},\boldsymbol{\psi}^{(2),*},\boldsymbol{\alpha}^*)$ has mean zero for all $\boldsymbol{\psi}^{(0)}$ and consequently its gradient w.r.t. $\boldsymbol{\psi}^{(0)}$ has mean zero under model $\mathcal{M}\cap\{\mathcal{M}(\boldsymbol{\alpha}^{(1)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(2)})\cap\mathcal{M}(\boldsymbol{\alpha}^{(0)})\}$. This leads to an unbiased estimating function for $\boldsymbol{\alpha}^{(0)}$ indexing the working model $\mathcal{M}(\boldsymbol{\alpha}^{(0)})$.

Because the gradients of $\phi_{\text{ext}}(\boldsymbol{O};\beta_0;\boldsymbol{\psi},\boldsymbol{\alpha})$ w.r.t. the nuisance parameters do not depend on the target parameter β_0, the result of Theorem 4.1 can be generalized to this multiply robust estimator to obtain the bias-reduced multiply robust estimator $\hat{\beta}_{n,\text{ext}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}},\hat{\boldsymbol{\alpha}}_n^{\text{BR}})$ with (the probability limits of) the nuisance parameter estimators $\hat{\boldsymbol{\psi}}_{n,\text{BR}}^{(0)}$, $\hat{\boldsymbol{\psi}}_{n,\text{BR}}^{(1)}$, $\hat{\boldsymbol{\psi}}_{n,\text{BR}}^{(2)}$, $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(0)}$, $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(1)}$ and $\hat{\boldsymbol{\alpha}}_{n,\text{BR}}^{(2)}$ minimizing the squared first-order bias of the multiply robust estimator.

4.6 Data Analysis: SUPPORT

We reanalyze data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) conducted in 1989-1994 in $n = 5735$ critically ill patients in five US hospitals to study the effectiveness of right heart catherization (RHC) in the initial care unit (ICU) of critically ill patients (Connors et al. 1996). RHC is a diagnostic procedure which, at the time of the study by Connors et al. (1996), was thought to lead to better patient outcomes by many physicians. The effectiveness of RHC had not been demonstrated in a randomized clinical trial but based on expert information, a rich set of 72 variables was collected to adjust for potential confounding (see Table 1 in Hirano and Imbens (2002)). The original analysis in Connors et al. (1996) used propensity score matching and surprisingly found that RHC leads to lower survival as compared to not performing RHC. For each patient, the treatment status A indicates 1 if RHC was applied within 24 hours of admission and 0 otherwise. In total, 2184 patients received RHC and 3551 did not. We consider the effect of RHC on 30-day survival Y with $Y = 1$ indicating survival, 0 otherwise. In total, 3817 patients survived and 1918 died within 30 days. The two treatment groups differ significantly in terms of the distributions of the 72 covariates $\boldsymbol{X}$. Figure 4.5 visualizes the large differences that exist in baseline covariate means between treated and untreated patients (see the x-axis). A detailed description is given in Table 2 of Hirano and Imbens (2002).

Estimators

To estimate the additive treatment effect $\tau = E\{Y(1)\} - E\{Y(0)\}$, we use the results of Section 4.5.1. As in Hirano and Imbens (2002), we model the propensity score $P(A = 1|\boldsymbol{X})$ using a logistic regression model including a constant term and all 72 main effects; $\pi(\boldsymbol{X};\boldsymbol{\psi}) = \text{expit}\{\boldsymbol{\psi}^T\boldsymbol{k}(\boldsymbol{X})\}$ with $\boldsymbol{k}(\boldsymbol{X}) = (1, X_1, \ldots, X_{72})$. We model the conditional mean outcome $E(Y|A = a, \boldsymbol{X})$ for $a \in \{0,1\}$ using both a linear and logistic regression model $m^{(a)}_{\text{lin}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)}) = \boldsymbol{\alpha}^{(a),T}\boldsymbol{k}(\boldsymbol{X})$ and $m^{(a)}_{\text{logit}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)}) = \text{expit}\{\boldsymbol{\alpha}^{(a),T}\boldsymbol{k}(\boldsymbol{X})\}$ for $a \in \{0,1\}$, thus both including a constant term and all 72 main effects.

For the linear outcome model, we obtain estimators for the nuisance parameters $(\hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR,lin}}, \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR,lin}}, \hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR,lin}}, \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR,lin}})$ solving estimating equations (4.18) and

Figure 4.5: *Comparison of unweighted and weighted covariate means. The weighted covariate means are calculated based on both an analysis using a linear outcome model (top) and a logit outcome model (bottom), see the Estimators-paragraph for more details.*

(4.19) for condition $A = 1$ and estimating equations (4.20) and (4.21) for condition $A = 0$ with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}) = \{1-\pi(\boldsymbol{X};\boldsymbol{\psi})\}\pi(\boldsymbol{X};\boldsymbol{\psi})\boldsymbol{k}(\boldsymbol{X})$ and $m^{(a)}_{\boldsymbol{\alpha}^{(a)},\text{lin}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)}) = \boldsymbol{k}(\boldsymbol{X})$. The estimator $\hat{\boldsymbol{\psi}}^{(a)}_{n,\text{BR,lin}}$ is obtained by maximizing the function

$$\mathcal{F}^{(a)}_{\text{lin}}(\boldsymbol{\psi}) = n^{-1}\sum_{i=1}^{n}\left[(-1)^a(A_i-1+a)\exp\left\{(-1)^a\boldsymbol{\psi}^T\boldsymbol{k}(\boldsymbol{X}_i)\right\} + (A_i-a)\boldsymbol{\psi}^T\boldsymbol{k}(\boldsymbol{X}_i)\right],$$

which is an integrated form of (4.18) for $a = 1$ and (4.20) for $a = 0$ in the sense that $\partial\mathcal{F}^{(a)}_{\text{lin}}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ equals (4.18) for $a = 1$ and (4.20) for $a = 0$, both using the linear outcome model $m^{(a)}_{\text{lin}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)}) = \boldsymbol{\alpha}^{(a),T}\boldsymbol{k}(\boldsymbol{X})$ ($a \in \{0,1\}$).

For the logistic outcome model, we obtain estimators for the nuisance parameters $(\hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR,logit}}, \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR,logit}}, \hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR,logit}}, \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR,logit}})$ also solving estimating equations (4.18) and (4.19) for condition $A = 1$ and estimating equations (4.20) and (4.21) for condition $A = 0$ with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}) = \{1-\pi(\boldsymbol{X};\boldsymbol{\psi})\}\pi(\boldsymbol{X};\boldsymbol{\psi})\boldsymbol{k}(\boldsymbol{X})$ but now with $m^{(a)}_{\boldsymbol{\alpha}^{(a)},\text{logit}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)}) = \{1-m^{(a)}_{\text{logit}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)})\}m^{(a)}_{\text{logit}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)})\boldsymbol{k}(\boldsymbol{X})$. The estimator $\hat{\boldsymbol{\psi}}^{(a)}_{n,\text{BR,logit}}$ is obtained by maximizing the function

$$\mathcal{F}^{(a)}_{\text{logit}}(\boldsymbol{\psi}) = n^{-1}\sum_{i=1}^{n}\Big\{\left[(-1)^a(A_i-1+a)\exp\left\{(-1)^a\boldsymbol{\psi}^T\boldsymbol{k}(\boldsymbol{X}_i)\right\} + (A_i-a)\boldsymbol{\psi}^T\boldsymbol{k}(\boldsymbol{X}_i)\right] \{1-m^{(a)}_{\text{logit}}(\boldsymbol{X}_i;\boldsymbol{\alpha}^{(a)})\}m^{(a)}_{\text{logit}}(\boldsymbol{X}_i;\boldsymbol{\alpha}^{(a)})\Big\},$$

which is an integrated form of (4.18) for $a = 1$ and (4.20) for $a = 0$ in the sense that $\partial\mathcal{F}^{(a)}_{\text{logit}}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ equals (4.18) for $a = 1$ and (4.20) for $a = 0$, both using the logistic outcome model $m^{(a)}_{\text{logit}}(\boldsymbol{X};\boldsymbol{\alpha}^{(a)}) = \text{expit}\{\boldsymbol{\alpha}^{(a),T}\boldsymbol{k}(\boldsymbol{X})\}$ ($a \in \{0,1\}$). Because $\mathcal{F}^{(a)}_{\text{logit}}(\boldsymbol{\psi})$ depends on the parameter $\boldsymbol{\alpha}^{(a)}$, we need to plug-in an initial estimator for this parameter in order to obtain $\hat{\boldsymbol{\psi}}^{(a)}_{n,\text{BR,logit}}$. Here, the function $\mathcal{F}^{(a)}_{\text{logit}}(\boldsymbol{\psi})$ is maximized using the MLE for $\boldsymbol{\alpha}^{(a)}$ ($a \in \{0,1\}$).

Doubly robust estimators for the additive treatment effect τ are then obtained as

$$\hat{\tau}_{n,\text{BR,lin}} = \hat{\mu}^{(1)}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR,lin}}, \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR,lin}}) - \hat{\mu}^{(0)}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR,lin}}, \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR,lin}}),$$

$$\hat{\tau}_{n,\text{BR,logit}} = \hat{\mu}^{(1)}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}^{(1)}_{n,\text{BR,logit}}, \hat{\boldsymbol{\alpha}}^{(1)}_{n,\text{BR,logit}}) - \hat{\mu}^{(0)}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}^{(0)}_{n,\text{BR,logit}}, \hat{\boldsymbol{\alpha}}^{(0)}_{n,\text{BR,logit}}).$$

The estimators of the propensity score are different, albeit similar, when estimating $E\{Y(1)\}$ versus $E\{Y(0)\}$. They reveal sufficient overlap of the propensity score distributions in the RHC group and the no-RHC group (see Figure 4.6). Figure 4.5 shows the unweighted t-statistics of standard t-tests comparing the group-specific covariate means (x-axis) versus the weighted t-statistics of weighted t-tests comparing the group-specific covariate means (y-axis), that is, each individual is weighted w.r.t. the estimated probability of getting the observed treatment. Results using weights based on $\hat{\boldsymbol{\psi}}^{(a)}_{n,\text{BR,lin}}$ (w.r.t. a linear outcome model) are displayed on top and results using weights based on $\hat{\boldsymbol{\psi}}^{(a)}_{n,\text{BR,logit}}$ (w.r.t. a logit outcome model) are shown below. It demonstrates that inverse probability of treatment weighting balances the RHC group and the no-RHC group very well.

Results

Below we summarize the data analysis results. We obtain an unadjusted effect estimate $\hat{\tau}_{n,\text{unadj}} = -0.0736$ (SE $= 0.0272$, 95% CI -0.1269 to -0.0203) which is prone to potential confounding. The standard doubly robust estimate for the average treatment effect using MLE for all working models equals $\hat{\tau}_{n,\text{MLE,lin}} = -0.0649$ (SE $= 0.0162$, 95% CI -0.0966 to -0.0332) and $\hat{\tau}_{n,\text{MLE,logit}} = -0.0657$ (SE $= 0.0158$, 95% CI -0.0967 to -0.0346). The biased-reduced doubly robust estimation procedure gives more efficient results: we obtain $\hat{\tau}_{n,\text{BR,lin}} = -0.0612$ (SE $= 0.0141$, 95% CI -0.0889 to -0.0335 and $\text{avar}(\hat{\tau}_{n,\text{MLE,lin}})/\text{avar}(\hat{\tau}_{n,\text{BR,lin}}) = 1.32$) and $\hat{\tau}_{n,\text{BR,logit}} = -0.0610$ (SE $= 0.0137$, 95% CI -0.0879 to -0.0340 and $\text{avar}(\hat{\tau}_{n,\text{MLE,logit}})/\text{avar}(\hat{\tau}_{n,\text{BR,logit}}) = 1.33$). Results for the other improved doubly robust estimators are similar, but less efficient. For example, the calibrated likelihood estimator of Tan gives $\hat{\tau}_{n,\text{TAN}} = -0.0622$ (SE $= 0.0154$, 95% CI -0.0924 to -0.0319) and the TMLE with default super-learner gives $\hat{\tau}_{n,\text{TMLE-SL}} = -0.0586$ (SE $= 0.0149$, 95% CI -0.0877 to -0.0295). Over the different doubly robust methods, the estimates of $E\{Y(1)\}$ range from 0.630 to 0.634 and the estimates of $E\{Y(0)\}$ vary from 0.687 to 0.696.

The results found here are very similar to those found in Hirano and Imbens

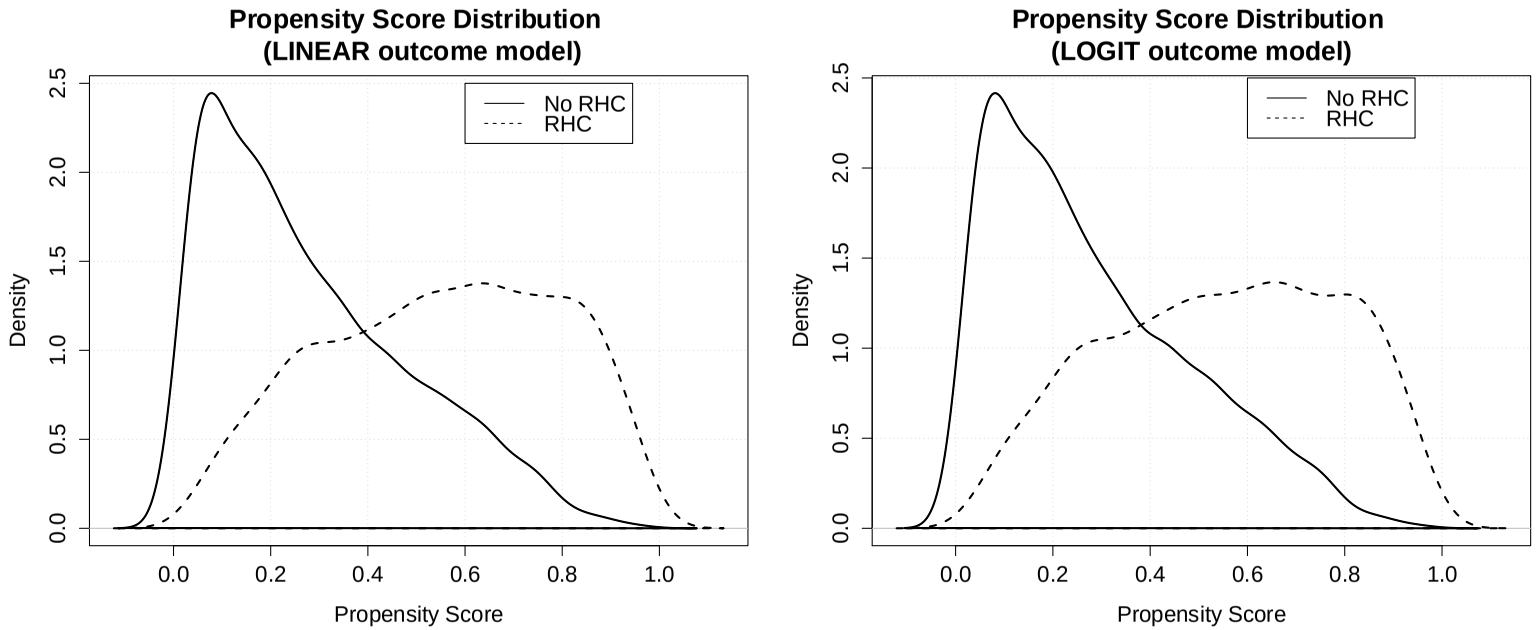


Figure 4.6: *Propensity score distribution based on the proposed estimators* $\hat{\boldsymbol{\psi}}^{(a)}_{n,BR,lin}$ *w.r.t. a linear outcome model (left) or* $\hat{\boldsymbol{\psi}}^{(a)}_{n,BR,logit}$ *w.r.t. a logit outcome model (right) for* $a = 0$*, No RHC (solid) and* $a = 1$*, RHC (dashed).*

(2002) and coincide with the initial findings from Connors et al. (1996). After adjustment for potential confounding and assuming no unmeasured confounding, RHC appears to lead to an increased 30 day mortality risk of critically ill patients in the initial care unit.

4.7 Discussion

In this chapter, we have proposed a novel strategy for estimating the nuisance parameters indexing the working models in doubly robust estimators. A defining property of the proposed bias-reduced doubly robust estimation strategy is that it locally minimizes the squared first-order asymptotic bias of the doubly robust estimator defined by finite-dimensional nuisance working models. It also makes the doubly robust estimator insensitive to local (one over root-n) perturbations of the nuisance parameters. This gets for instance reflected in improved stability of the weights in those doubly robust estimators that invoke inverse weighting. A corresponding efficiency benefit is hence logically anticipated. Formalizing this is however complicated by the fact that the choice of root-n consistent estimators for the nuisance parameters affects the asymptotic distribution of the doubly robust estimator not only through their own asymptotic distribution, but also through their probability limits, which can be different for each choice of estimator under model misspecification. In future work, we hope to develop further insight into the theoretical properties of bias-reduced doubly robust estimators as well as confidence intervals obtained by inverting score tests based on this strategy.

The principle of the bias-reduced doubly robust estimator is easy to use and adapts to a wide variety of doubly robust estimators. In that sense, it differs from the various other targeted proposals that have been made over recent years (Cao et al. 2009; Tan 2010; van der Laan and Gruber 2010; Tsiatis et al. 2011; van der Laan and Rose 2011; Rotnitzky et al. 2012; van der Laan 2014), some of which are not straightforward or even impossible to adapt to general doubly robust estimators (for instance when the observed data likelihood does not factorize). The simplicity of the proposed approach not only comes through the fact that the estimating functions for the nuisance parameters are readily obtained as gradients of the doubly robust

influence function under fixed nuisance parameters, but also through the fact that the asymptotic variance calculation of the resulting doubly robust estimator (and corresponding score tests) can ignore estimation of the nuisance parameters.

While the proposed approach is expected to yield estimators with reasonable precision, it does not guarantee minimal variance, unlike other proposals in certain settings (Cao et al. 2009). However, the requirement of minimal variance generally leads to complex constrained optimization problems (unless for instance when the propensity score model is assumed to be correctly specified). For instance, in the specific missing data problem that we considered in Section 4.3, Cao et al. (2009) showed that when the working model $\mathcal{M}(\boldsymbol{\psi})$ for the missingness probabilities is correctly specified, the parameters indexing the outcome working model can be estimated by minimizing the variance of the doubly robust estimator. This same principle does not work for estimating the parameters indexing the model for the missingness probabilities when instead the outcome working model $\mathcal{M}(\boldsymbol{\xi})$ is assumed to be correctly specified. To see this, we follow a similar reasoning as in Cao et al. (2009). For simplicity, consider the case where the working model $m(\boldsymbol{X})$ for the conditional mean outcome $m_0(\boldsymbol{X})$ is completely specified (i.e., known). The asymptotic variance of the doubly robust estimator

$$\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi}) = n^{-1}\sum_{i=1}^{n}\left\{\frac{R_i Y_i}{\pi(\boldsymbol{X}_i,\boldsymbol{\psi})} - \frac{R_i-\pi(\boldsymbol{X}_i,\boldsymbol{\psi})}{\pi(\boldsymbol{X}_i,\boldsymbol{\psi})}m(\boldsymbol{X}_i)\right\},$$

under a correctly specified model $m(\boldsymbol{X})$ for the conditional mean outcome $m_0(\boldsymbol{X})$ and a possibly misspecified working model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for the propensity score $\pi_0(\boldsymbol{X})$ is given by

$$\mathrm{var}\{m_0(\boldsymbol{X})\} + E\left\{\frac{\pi_0(\boldsymbol{X})}{\pi^2(\boldsymbol{X};\boldsymbol{\psi})}V(\boldsymbol{X})\right\},$$

with $V(\boldsymbol{X}) = \mathrm{var}(Y|\boldsymbol{X})$. We assume homoscedasticity such that $V(\boldsymbol{X}) = \sigma^2$. Because we want to minimize the asymptotic variance of the doubly robust estimator at $\mathcal{M}(\boldsymbol{\xi})$, one might define $\boldsymbol{\psi}^*$ to be the value that minimizes this variance. However, taking the gradient w.r.t. $\boldsymbol{\psi}$ of the asymptotic variance does not automatically lead to an unbiased estimating function at $\mathcal{M}(\boldsymbol{\psi})$. This implies that $\boldsymbol{\psi}^*$ will not necessarily equal the truth when $\mathcal{M}(\boldsymbol{\psi})$ holds. Instead, $\boldsymbol{\psi}^*$ must therefore be defined as $\boldsymbol{\psi}_{\boldsymbol{d}}^* = \mathrm{argmin}_{\boldsymbol{\psi}_{\boldsymbol{d}}}\left[E\left\{R/\pi^2(\boldsymbol{X},\boldsymbol{\psi}_{\boldsymbol{d}})\right\}\right]$ where $\boldsymbol{\psi}_{\boldsymbol{d}}$ must satisfy

$E\left[\boldsymbol{d}(\boldsymbol{X})\left\{R/\pi(\boldsymbol{X},\boldsymbol{\psi}_{\boldsymbol{d}})-1\right\}\right]=\mathbf{0}$ for some function $\boldsymbol{d}(\boldsymbol{X})$ of the dimension of $\boldsymbol{\psi}_{\boldsymbol{d}}$. This $\boldsymbol{\psi}_{\boldsymbol{d}}^{*}$ has the property of minimizing the asymptotic variance of $\hat{\mu}_{n,\mathrm{DR}}(\boldsymbol{\psi})$ when $m(\boldsymbol{X})$ is correctly specified under the condition it leads to an unbiased estimating function and equals the truth when $\pi(\boldsymbol{X};\boldsymbol{\psi})$ is correctly specified. An estimator $\hat{\boldsymbol{\psi}}_{n,\boldsymbol{d}}$ for $\boldsymbol{\psi}_{\boldsymbol{d}}^{*}$ is then obtained by using the empirical analog of the constrained optimization problem.

The bias-reduced doubly robust estimation principle may more generally lend itself better to small-sample inference. For instance, suppose that interest lies in the marginal causal effect $\tau=E\{Y(1)-Y(0)\}$. Because the proposed estimation strategy does not require acknowledging the uncertainty of the estimated nuisance parameters (up to first order), we foresee that it may potentially lend itself better to randomization inference (e.g., permutation tests). How such randomization inference could be accomplished and how it performs in small to large samples will be studied in future work.

A limitation of the bias-reduced doubly robust estimator is that it demands working models of the same dimension. This can in principle be remedied by enlarging the working models with clever choices of covariates until they are of the same dimension. For example, reconsider the doubly robust estimator (3.3) from Section 3.3 with working models $m(X;\boldsymbol{\xi})=\xi_1+\xi_2X+\xi_3X^2$ and $\pi(X;\boldsymbol{\psi})=\operatorname{expit}(\psi_1+\psi_2X)$ for a one-dimensional covariate X. Taking the gradients of the influence function would lead to two estimating functions for $\boldsymbol{\xi}=(\xi_1,\xi_2,\xi_3)^T$ and three estimating functions for $\boldsymbol{\psi}=(\psi_1,\psi_2)^T$. An additional estimating function for $\boldsymbol{\xi}$ can be obtained by using the extended propensity score model $\pi(X;\boldsymbol{\psi})=\operatorname{expit}\{\psi_1+\psi_2X+\psi_3\zeta(X)\}$ for a cleverly chosen covariate $\zeta(X)$. The proposal then amounts to solving the estimating equations

$$\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{n}\partial\phi(\boldsymbol{O}_i;\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})/\partial\boldsymbol{\xi} \\
&= \sum_{i=1}^{n}\left\{1-\frac{R_i}{\pi(X_i;\boldsymbol{\psi})}\right\}\begin{bmatrix}1\\X_i\\X_i^2\end{bmatrix} \\
\mathbf{0} &= \sum_{i=1}^{n}\partial\phi(\boldsymbol{O}_i;\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})/\partial\boldsymbol{\psi}
\end{aligned}$$

$$= \sum_{i=1}^{n} A_i \left\{ \frac{1-\pi(X_i;\boldsymbol{\psi})}{\pi(X_i;\boldsymbol{\psi})} \right\} \{Y_i - m(X_i;\boldsymbol{\xi})\} \begin{bmatrix} 1 \\ X_i \\ \zeta(X_i) \end{bmatrix}$$

for $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, respectively. A clever choice for $\zeta(X)$ is $1/\{1-\pi(X;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})\}$. Indeed, this choice ensures that

$$\sum_{i=1}^{n} \frac{A_i}{\pi(X_i;\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}})} \{Y_i - m(X_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})\} = 0,$$

making the bias-reduced doubly robust estimator $\hat{\mu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n^{\mathrm{BR}}, \hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$ equal to a substitution estimator $n^{-1}\sum_{i=1}^{n} m(X_i;\hat{\boldsymbol{\xi}}_n^{\mathrm{BR}})$. An alternative possibility to cope with nuisance parameters of different dimensions would be to apply the procedure in the direction of just a single nuisance parameter, rather than both, which we will explore in the next chapter.

4.A Regularity Conditions

Interchanging integration and differentiation

Let $\mathcal{B}$ be an open subset in $\mathbb{R}^r$. Under the conditions (i) $\phi(\boldsymbol{o};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi})$ is integrable as a function of $\boldsymbol{o}$ (w.r.t. the probability measure $F_0(\boldsymbol{o})$) for all $\boldsymbol{\xi}\in\mathcal{B}$, (ii) for all j, with probability one, the derivatives $\partial\phi(\boldsymbol{o};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi})/\partial\xi_j$ exist for all $\boldsymbol{\xi}\in\mathcal{B}$ where ξ_j is the jth component of $\boldsymbol{\xi}$ and (iii) for all j, there is an integrable function $B_j(\boldsymbol{o})$ such that $|\partial\phi(\boldsymbol{o};\mu_0,\psi_0,\boldsymbol{\xi})/\partial\xi_j|\leq B_j(\boldsymbol{o})$ for all $\boldsymbol{\xi}\in\mathcal{B}$, we have for all j that

$$\frac{\partial}{\partial\xi_j}\int\phi(\boldsymbol{o};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi})\mathrm{d}F_0(\boldsymbol{o})=\int\frac{\partial}{\partial\xi_j}\phi(\boldsymbol{o};\mu_0,\boldsymbol{\psi}_0,\boldsymbol{\xi})\mathrm{d}F_0(\boldsymbol{o}).$$

Similar conditions need to hold with the role of $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ reversed.

4.B Bias of the Doubly Robust Estimator with Estimated Nuisance Parameters

Let $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\xi}}_n$ be arbitrary root-n consistent estimators for their corresponding probability limits $\boldsymbol{\psi}^*$ and $\boldsymbol{\xi}^*$. Let $\phi_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi})$ and $\phi_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\xi})$ be the corresponding influence functions in the sense that

$$(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}^*)=n^{-1}\sum_{i=1}^{n}\phi_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)+o_p(n^{-1/2})$$

and

$$(\hat{\boldsymbol{\xi}}_n-\boldsymbol{\xi}^*)=n^{-1}\sum_{i=1}^{n}\phi_{\boldsymbol{\xi}}(\boldsymbol{O}_i;\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\}+o_p(n^{-1/2}).$$

These influence functions are unbiased at the corresponding probability limits of the nuisance parameter estimators, $E\{\phi_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\}=\mathbf{0}$ and $E\{\phi_{\boldsymbol{\xi}}(\boldsymbol{O};\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)\}=\mathbf{0}$. Consider a standard second order Taylor expansion of $\mathrm{bias}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n;\mu_0)$ around the limiting values $\boldsymbol{\psi}^*$ and $\boldsymbol{\xi}^*$ (recall that $\mathrm{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)=E\{\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})\}$),

$$\begin{aligned}&\mathrm{bias}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n;\mu_0)\\&\quad=\mathrm{bias}(\boldsymbol{\psi}^*,\boldsymbol{\xi}^*;\mu_0)\end{aligned}$$

$$+ E\left\{ \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\psi}^T}\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*) & \frac{\partial}{\partial \boldsymbol{\xi}^T}\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^* \\ \hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^* \end{bmatrix} \right\}$$
$$+ \frac{1}{2} E\left\{ \begin{bmatrix} \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^* \\ \hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^* \end{bmatrix}^T \boldsymbol{H}(\boldsymbol{O};\tilde{\boldsymbol{\psi}}_n,\tilde{\boldsymbol{\xi}}_n) \begin{bmatrix} \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^* \\ \hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^* \end{bmatrix} \right\}$$

for intermediate values $(\tilde{\boldsymbol{\psi}}_n^T, \tilde{\boldsymbol{\xi}}_n^T)^T$ on the line segment joining $(\hat{\boldsymbol{\psi}}_n^T, \hat{\boldsymbol{\xi}}_n^T)^T$ and $(\boldsymbol{\psi}^{*,T}, \boldsymbol{\xi}^{*,T})^T$ and

$$\boldsymbol{H}(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\xi}) = \begin{bmatrix} \partial^2\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T & \partial^2\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\xi}^T \\ \partial^2\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})/\partial\boldsymbol{\xi}\partial\boldsymbol{\psi}^T & \partial^2\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi},\boldsymbol{\xi})/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T \end{bmatrix}.$$

Assume that $\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)/\partial\boldsymbol{\xi}^T$, $\partial\phi(\boldsymbol{O};\mu_0,\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)/\partial\boldsymbol{\psi}^T$ and $\boldsymbol{H}(\boldsymbol{O};\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)$ are $O_p(1)$. Next, assuming $\boldsymbol{H}$ is a continuous function of the nuisance parameters, $\boldsymbol{H}(\boldsymbol{O};\tilde{\boldsymbol{\psi}}_n,\tilde{\boldsymbol{\xi}}_n) \xrightarrow{p} \boldsymbol{H}(\boldsymbol{O};\boldsymbol{\psi}^*,\boldsymbol{\xi}^*)$ and because of this convergence in probability, $\boldsymbol{H}(\boldsymbol{O};\tilde{\boldsymbol{\psi}}_n,\tilde{\boldsymbol{\xi}}_n)$ is also $O_p(1)$. From the root-n consistency of the nuisance parameter estimators, we can conclude that

$$\text{bias}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\xi}}_n;\mu_0) = \text{bias}(\boldsymbol{\psi}^*,\boldsymbol{\xi}^*) + o(1).$$

4.C R-Functions

Below, we provide R-functions to obtain the bias-reduced doubly robust estimator $\hat{\mu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}}, \hat{\boldsymbol{\xi}}_n^{\text{BR}})$ to estimate the mean outcome $E(Y)$ in the presence of incomplete data as outlined in Section 4.3 for both a linear regression working model and logistic regression working model for the conditional mean outcome.

As input, both functions use the missingness indicator `R`, the outcome `Y` and the auxiliary covariates `cov`. As output, they deliver the bias-reduced doubly robust estimate `mn.Y` of the mean outcome, the corresponding standard error `se.mn.Y`, the estimates `psi.BR` of the parameters indexing the working model for the missingness probabilities and the estimates `xi.BR` of the parameters indexing the working model for the conditional mean.

Linear outcome working model

```
m.biasreducedDR.identity<-function(R,Y,cov){
  n<-length(R)
  int.cov<-cbind(rep(1,n),cov)
  expit<-function(x) exp(x)/(1+exp(x))
  U <- function(R,Y,X,psi,xi){
    (R/expit(psi%*%t(X))*(Y-xi%*%t(X))+xi%*%t(X))
  }
  min.Uint<-function(psi){
    -mean((-R*exp(-psi%*%t(int.cov)))+(-(1-R)
                  *(psi%*%t(int.cov))))
  }

  init.psi<-coef(glm(R~cov,family="binomial"))
  sol<-nlm(min.Uint,init.psi)
  psi.BR<-sol$estimate
  weight<-as.vector(1/exp(psi.BR%*%t(int.cov)))
  xi.BR<-coef(lm(Y~-1+int.cov,subset=(R==1),weights=weight))

  mn.Y<-mean(U(R,Y,int.cov,psi.BR,xi.BR))
  se.mn.Y<-sd(U(R,Y,int.cov,psi.BR,xi.BR))/sqrt(n)

  return(list(mn.Y=mn.Y,se.mn.Y=se.mn.Y,
                    psi.BR=psi.BR,xi.BR=xi.BR))
}
```

Logistic outcome working model

```
m.biasreducedDR.logit<-function(R,Y,cov){
  n<-length(R)
  int.cov<-cbind(rep(1,n),cov)
  expit<-function(x) exp(x)/(1+exp(x))
  U <- function(R,Y,X,psi,xi){
    (R/expit(psi%*%t(X))*(Y-xi%*%t(X))+xi%*%t(X))
  }
  min.Uint<-function(psi){
    -mean(((-R*exp(-osi%*%t(int.cov)))+(-(1-R)
                    *(psi%*%t(int.cov))))
             *as.vector(expit(init.xi%*%t(int.cov))
             *(1-expit(init.xi%*%t(int.cov)))))
  }

  init.xi<-coef(glm(R~cov,family="binomial"))
  init.psi<-coef(glm(Y~cov,family="binomial",subset=(R==1)))
```

```
  sol<-nlm(min.Uint,init.psi)
  psi.BR<-sol$estimate
  weight<-as.vector(1/exp(psi.BR%*%t(int.cov)))
  xi.BR<-coef(glm(Y~-1+int.cov,family="binomial",
              subset=(R==1),weights=weight))

  mn.Y<-mean(U(R,Y,int.cov,psi.BR,xi.BR))
  se.mn.Y<-sd(U(R,Y,int.cov,psi.BR,xi.BR))/sqrt(n)

  return(list(mn.Y=mn.Y,se.mn.Y=se.mn.Y,
                    psi.BR=psi.BR,xi.BR=xi.BR))
}
```

CHAPTER 5

Data-Adaptive Bias-Reduced Doubly Robust Estimation

In Chapter 3, we demonstrated that doubly robust estimators consistently estimate the parameter of interest in large semiparametric models when one of two nuisance working models is correctly specified, regardless of which. In Chapter 4 (see also Vermeulen and Vansteelandt (2015a)), we expanded this robustness property to more realistic settings where both working models are misspecified. In particular, the **bias-reduced doubly robust estimators** make use of special nuisance parameter estimators that are designed to locally minimize the squared first-order asymptotic bias of the doubly robust estimator under misspecification of both working models. In this chapter, we extend this idea to incorporate the use of data-adaptive estimators in an attempt to further reduce bias. Simulation studies confirm the desirable finite-sample performance of the proposed estimators relative to a variety of other doubly robust estimators. This chapter is based on Vermeulen and Vansteelandt (2015b).

5.1 Introduction

In Chapter 4, we investigated the usefulness of doubly robust estimators from the perspective that **all working models are wrong**. We discovered that, interestingly, some doubly robust estimators partially retain their robustness properties, even under misspecification of both working models where we in particular studied the bias of doubly robust estimators for a scalar parameter. We found that without knowledge of the true data-generating law, the bias of these estimators can be locally minimized across all values of the nuisance parameters indexing parametric working models. This is possible by making use of specific estimators of the nuisance parameters, which target bias reduction. We referred to this procedure as **bias-reduced doubly robust estimation**.

In Section 4.4, we contrasted the bias-reduced doubly robust estimator with a variety of other doubly robust estimators that are primarily aimed at variance reduction under misspecification of one working model. We found the bias-reduced doubly robust estimator to be highly competitive, although sometimes somewhat more biased than so-called targeted maximum likelihood estimators (TMLE) (van der Laan and Rubin 2006). These reduce bias by making clever use of data-adaptive learning algorithms such as ensemble learning, and specifically super-learning (van der Laan et al. 2007). In this chapter, we will investigate how such data-adaptive learning algorithms can be integrated in the bias-reduced doubly robust estimation procedure to allow for further bias reduction. This will also overcome one of the limitations of the bias-reduced estimation procedure, that both nuisance working models must be indexed by nuisance parameters of the same dimension.

For convenience, we introduce some new notation in Section 5.2 to remain close to the existing TMLE-literature and rephrase the inferential problem of the estimation of a population mean outcome in the presence of missingness that is explainable by measured auxiliary covariates using this notation. In Section 5.3, we briefly review the theory on biased-reduced doubly robust estimation in the context of the example given in Section 5.2 in this new notation. Next, in Section 5.4, we outline the proposed extension of the bias-reduced doubly robust estimator to incorporate data-adaptive learning algorithms. In addition, we show how to perform inference based on the asymptotic linearity of the estimator (under certain regularity

conditions, given in Appendix 5.A). In Section 5.5, we illustrate the performance of our proposal relative to the original bias-reduced doubly robust estimator and the TMLE procedure (van der Laan and Rubin 2006). In Section 5.6, we illustrate the generic nature of the proposal and show how to implement the proposed strategy in a linear instrumental variable analysis. We end with a discussion in Section 5.7.

5.2 Doubly Robust Estimation of a Population Mean With Incomplete Data

Consider the i.i.d. sample $\mathcal{O} = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n)$ of size n, where the observed data vector $\boldsymbol{O} = (RY, R, \boldsymbol{X})$ is distributed according to a true underlying but unknown probability distribution P_0. As before, we let Y denote the outcome of interest which is susceptible to missingness, formalized using the missingness indicator R, where $R = 1$ if Y is observed and $R = 0$ if Y is missing. We assume that missingness is explainable by $\boldsymbol{X}$, a collection of auxiliary covariates, so that the missing at random (MAR, Rubin (1976)) assumption holds: $Y \perp\!\!\!\perp R|\boldsymbol{X}$. Throughout, for a function f of the observed data and a probability distribution P, we let Pf denote the integral $\int f dP$.

Doubly robust estimation of μ_0 requires specification of two nuisance working models (Scharfstein et al. 1999a). First, we need a working model

$$\mathcal{G} = \{g(\boldsymbol{X})|g \text{ in some class of functions}\}$$

for the true missingness mechanism $g_0(\boldsymbol{X}) = P_0(R = 1|\boldsymbol{X})$, referred to as the propensity score, for which we assume **positivity**: $g_0(\boldsymbol{X}) \geq \delta > 0$ with probability one (van der Laan and Rose 2011, chap. 10). Let $\mathcal{M}(\mathcal{G})$ denote the statistical model for the joint distribution of $\boldsymbol{O}$ implied by the working model $\mathcal{G}$. Let $\hat{g}_n(\boldsymbol{X})$ denote an estimator of $g_0(\boldsymbol{X})$ with probability limit $g^*(\boldsymbol{X})$; that is, $\hat{g}_n(\boldsymbol{X}) \xrightarrow{p} g^*(\boldsymbol{X})$. Under $\mathcal{M}(\mathcal{G})$, $g^*(\boldsymbol{X}) = g_0(\boldsymbol{X})$. Second, we need a working model

$$\mathcal{Q} = \{\bar{Q}(\boldsymbol{X})|\bar{Q} \text{ in some class of functions}\}$$

for the true conditional mean outcome $\bar{Q}_0(\boldsymbol{X}) = E_0(Y|\boldsymbol{X})$ (which equals $E_0(Y|R =$

$1, \boldsymbol{X}$) because of MAR). Let $\mathcal{M}(\mathcal{Q})$ denote the statistical model for the joint distribution of $\boldsymbol{O}$ implied by $\mathcal{Q}$. Let $\hat{\bar{Q}}_n(\boldsymbol{X})$ denote an estimator of $\bar{Q}_0(\boldsymbol{X})$ with probability limit $\bar{Q}^*(\boldsymbol{X})$; that is, $\hat{\bar{Q}}_n(\boldsymbol{X}) \xrightarrow{p} \bar{Q}^*(\boldsymbol{X})$. Under $\mathcal{M}(\mathcal{Q})$, $\bar{Q}^*(\boldsymbol{X}) = \bar{Q}_0(\boldsymbol{X})$, but not otherwise. A doubly robust estimator of μ_0 is then obtained via

$$\hat{\mu}_{n,\mathrm{DR}}(\hat{g}_n, \hat{\bar{Q}}_n) = n^{-1} \sum_{i=1}^{n} \left[\frac{R_i}{\hat{g}_n(\boldsymbol{X}_i)} \{Y_i - \hat{\bar{Q}}_n(\boldsymbol{X}_i)\} + \hat{\bar{Q}}_n(\boldsymbol{X}_i) \right], \tag{5.1}$$

see also equations (3.3) and (3.4). This estimator is consistent for μ_0 under the union model $\mathcal{M}(\mathcal{G}) \cup \mathcal{M}(\mathcal{Q})$: as soon as one but not necessarily both working models is correctly specified. Provided sufficient regularity for the working models, it is also locally efficient (Bickel et al. 1993b) at the intersection model $\mathcal{M}(\mathcal{G}) \cap \mathcal{M}(\mathcal{Q})$ in the following sense: it has smallest asymptotic variance within the class of all estimators that are consistent and asymptotically normal under $\mathcal{M}(\mathcal{G})$, provided that also $\mathcal{M}(\mathcal{Q})$ is correctly specified. At the intersection model, it has the following simple expansion

$$\hat{\mu}_{n,\mathrm{DR}}(\hat{g}_n, \hat{\bar{Q}}_n) - \psi_0 = (P_0 - P_n)\{D^*(g_0, \bar{Q}_0; \mu_0)\} + R_n, \tag{5.2}$$

with the remainder $R_n = o_p(n^{-1/2})$ and P_n the empirical distribution which puts mass $1/n$ on each observation $\boldsymbol{O}_i$, $i = 1, \ldots, n$ and $D^*(g_0, \bar{Q}_0; \mu_0)$ is the efficient influence function, given by

$$D^*(g_0, \bar{Q}_0; \mu_0)(\boldsymbol{O}) = \frac{R}{g_0(\boldsymbol{X})} \{Y - \bar{Q}_0(\boldsymbol{X})\} + \bar{Q}_0(\boldsymbol{X}) - \mu_0. \tag{5.3}$$

This is attractive because the expansion (and thus the asymptotic distribution of the doubly robust estimator) is the same, no matter how the nuisance parameters are estimated and no matter whether they are estimated or known. Finally, note that by construction

$$P_n \left[D^* \left\{ \hat{g}_n, \hat{\bar{Q}}_n; \hat{\mu}_{n,\mathrm{DR}}(\hat{g}_n, \hat{\bar{Q}}_n) \right\} \right] = 0. \tag{5.4}$$

5.3 Bias-Reduced Doubly Robust Estimation

In Chapter 4, we demonstrated that bias-reduced doubly robust estimation is a generic estimation strategy for the nuisance parameters indexing parametric nuisance working models $\mathcal{G}$ and $\mathcal{Q}$, aimed at bias reduction under misspecification of both working models. In Section 5.3.1, we will introduce such parametric working models in the new notation and in Section 5.3.2, we will briefly review the bias-reduced estimation principle.

5.3.1 Parametric nuisance working models and MLE

Suppose we parameterize the working model for the propensity score by an s-dimensional parameter $\boldsymbol{\psi}$:

$$\mathcal{G}_{\boldsymbol{\psi}} = \{g(\boldsymbol{\psi})(\boldsymbol{X}) | \boldsymbol{\psi} \in \mathbb{R}^s\},$$

where $g(\boldsymbol{\psi})(\boldsymbol{X}) = G\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\}$, G an appropriate inverse link function and $\boldsymbol{l} = (1, l_1, \ldots, l_{s-1})$; e.g., a logistic regression model $g(\boldsymbol{\psi})(\boldsymbol{X}) = \text{expit}(\psi_1 + \boldsymbol{\psi}_2^T \boldsymbol{X})$, $\text{expit}(x) = 1/(1+e^{-x})$. If the model $\mathcal{M}(\mathcal{G}_{\boldsymbol{\psi}})$ is correctly specified and thus includes P_0, we let $\boldsymbol{\psi}_0$ be such that $g(\boldsymbol{\psi}_0) = g_0$. Further, let the r-dimensional parameter $\boldsymbol{\xi}$ parameterize the working model for the conditional mean outcome:

$$\mathcal{Q}_{\boldsymbol{\xi}} = \{\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) | \boldsymbol{\xi} \in \mathbb{R}^r\}$$

with $\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) = Q\{\boldsymbol{\xi}^T \boldsymbol{k}(\boldsymbol{X})\}$, Q an appropriate inverse link function and $\boldsymbol{k} = (1, k_1, \ldots, k_{r-1})$; e.g., a linear regression model $\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) = \xi_1 + \boldsymbol{\xi}_2^T \boldsymbol{X}$ for a continuous outcome Y. If the model $\mathcal{M}(\mathcal{Q}_{\boldsymbol{\xi}})$ is correctly specified and thus includes P_0, we let $\boldsymbol{\xi}_0$ be such that $\bar{Q}(\boldsymbol{\xi}_0) = \bar{Q}_0$.

Root-n consistent and asymptotically normal estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\xi}}_n$ for the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ can be obtained as solutions to estimating equations $P_n\left\{D_g(\hat{\boldsymbol{\psi}}_n)\right\} = \mathbf{0}$ and $P_n\left\{D_{\bar{Q}}(\hat{\boldsymbol{\xi}}_n)\right\} = \mathbf{0}$ with the estimating functions D_g and $D_{\bar{Q}}$ such that $P_0\{D_g(\boldsymbol{\psi}_0)\} = \mathbf{0}$ if $P_0 \in \mathcal{M}(\mathcal{G}_{\boldsymbol{\psi}})$ and $P_0\{D_{\bar{Q}}(\boldsymbol{\xi}_0)\} = \mathbf{0}$ if $P_0 \in \mathcal{M}(\mathcal{Q}_{\boldsymbol{\xi}})$. Throughout, we will assume that $\hat{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}^*$ (with $\boldsymbol{\psi}_0 = \boldsymbol{\psi}^*$ under model $\mathcal{M}(\mathcal{G}_{\boldsymbol{\psi}})$) and that $\hat{\boldsymbol{\xi}}_n \xrightarrow{p} \boldsymbol{\xi}^*$ (with $\boldsymbol{\xi}_0 = \boldsymbol{\xi}^*$ under model $\mathcal{M}(\mathcal{Q}_{\boldsymbol{\xi}})$). We more-

over assume that these estimators are asymptotically linear with influence function $-P_0\left\{D_{\boldsymbol{\psi},g}(\boldsymbol{\psi}^*)\right\}^{-1}D_g(\boldsymbol{\psi}^*)$ and $-P_0\{D_{\boldsymbol{\xi},\bar{Q}}(\boldsymbol{\xi}^*)\}^{-1}D_{\bar{Q}}(\boldsymbol{\xi}^*)$, with derivatives $D_{\boldsymbol{\psi},g}(\boldsymbol{\psi}^*)=\partial D_g(\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}$ and $D_{\boldsymbol{\xi},\bar{Q}}(\boldsymbol{\xi}^*)=\partial D_{\bar{Q}}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}|_{\xi=\boldsymbol{\xi}^*}$.

In practice, maximum likelihood estimation (MLE) (or least squares) is routinely employed to estimate the nuisance parameters. The MLEs $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{MLE}}$ solve the estimating equations $P_n\left\{D_g^{\text{MLE}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}})\right\}=\mathbf{0}$ and $P_n\left\{D_{\bar{Q}}^{\text{MLE}}(\hat{\boldsymbol{\xi}}_n^{\text{MLE}})\right\}=\mathbf{0}$, with

$$D_g^{\text{MLE}}(\boldsymbol{\psi})(\boldsymbol{O}) = \frac{R-g(\boldsymbol{\psi})(\boldsymbol{X})}{g(\boldsymbol{\psi})(\boldsymbol{X})\{1-g(\boldsymbol{\psi})(\boldsymbol{X})\}}G'\{\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X})\}\boldsymbol{l}(\boldsymbol{X}), \quad (5.5)$$

$$D_{\bar{Q}}^{\text{MLE}}(\boldsymbol{\xi})(\boldsymbol{O}) = R\{Y-\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X})\}Q'\{\boldsymbol{\xi}^T\boldsymbol{k}(\boldsymbol{X})\}\boldsymbol{k}(\boldsymbol{X}), \quad (5.6)$$

$G'(x)=\partial G(x)/\partial x$ and $Q'(x)=\partial Q(x)/\partial x$ (see also Section 2.5.3). The corresponding doubly robust estimator is then given by $\hat{\mu}_{n,\text{MLE}}\equiv\hat{\mu}_{n,\text{DR}}(\hat{g}_n^{\text{MLE}},\hat{\bar{Q}}_n^{\text{MLE}})$, with $\hat{g}_n^{\text{MLE}}\equiv g(\hat{\boldsymbol{\psi}}_n^{\text{MLE}})$ and $\hat{\bar{Q}}_n^{\text{MLE}}\equiv\bar{Q}(\hat{\boldsymbol{\xi}}_n^{\text{MLE}})$. Although the MLE is asymptotically efficient and optimal for these nuisance parameters with respect to the corresponding working model, it need not be optimal with respect to the target parameter μ_0 under misspecification of one of these models. Under such misspecification, the influence function of the doubly robust estimator (and thus its asymptotic distribution) becomes indeed dependent on the choice of root-n consistent estimators of the nuisance parameters (see Proposition 3.1). This raises the question how to best fit the nuisance working models.

5.3.2 Bias-reduced doubly robust estimation

In Chapter 4, we showed that nuisance parameter estimators can be constructed whose probability limits locally minimize the squared first-order asymptotic bias of the doubly robust estimator under misspecification of both working models over all nuisance parameter values. Interestingly, computation of these estimators does not demand additional assumptions on the full data law. Indeed, under possible misspecification of $\mathcal{G}_{\boldsymbol{\psi}}$ and $\mathcal{Q}_{\boldsymbol{\xi}}$, the first-order asymptotic bias of the doubly robust estimator, as a function of P_0 and the nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, is given by $\text{bias}(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)=P_0\left[D^*\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi});\mu_0\}\right]$. This bias is zero under the union model $\mathcal{M}(\mathcal{G}_{\boldsymbol{\psi}})\cup\mathcal{M}(\mathcal{Q}_{\boldsymbol{\xi}})$. Away from this model, the squared first-

order bias bias$^2(\boldsymbol{\psi},\boldsymbol{\xi};\mu_0)$ is locally minimized in the direction of $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ at the values $(\boldsymbol{\psi}^{*,T}_{\text{BR}},\boldsymbol{\xi}^{*,T}_{\text{BR}})^T$ that solve the equations $P_0\left[D^*_{\boldsymbol{\psi}}\{g(\boldsymbol{\psi}^*_{\text{BR}}),\bar{Q}(\boldsymbol{\xi}^*_{\text{BR}})\}\right]=\mathbf{0}$ and $P_0\left[D^*_{\boldsymbol{\xi}}\{g(\boldsymbol{\psi}^*_{\text{BR}}),\bar{Q}(\boldsymbol{\xi}^*_{\text{BR}})\}\right]=\mathbf{0}$ (see Theorem 4.1); here, $D^*_{\boldsymbol{\psi}}\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi})\}$ denotes the gradient $\partial D^*\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi});\mu_0\}/\partial\boldsymbol{\psi}$ and $D^*_{\boldsymbol{\xi}}\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi})\}$ denotes the gradient $\partial D^*\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi});\mu_0\}/\partial\boldsymbol{\xi}$. These unknown population values can then be estimated via estimators $\hat{\boldsymbol{\psi}}^{\text{BR}}_n$ and $\hat{\boldsymbol{\xi}}^{\text{BR}}_n$ that solve the estimating equations

$$P_n\left[D^*_{\boldsymbol{\xi}}\{g(\hat{\boldsymbol{\psi}}^{\text{BR}}_n),\bar{Q}(\hat{\boldsymbol{\xi}}^{\text{BR}}_n)\}\right] = \mathbf{0}, \tag{5.7}$$

$$P_n\left[D^*_{\boldsymbol{\psi}}\{g(\hat{\boldsymbol{\psi}}^{\text{BR}}_n),\bar{Q}(\hat{\boldsymbol{\xi}}^{\text{BR}}_n)\}\right] = \mathbf{0}. \tag{5.8}$$

Here, the function $D^*_{\boldsymbol{\xi}}\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi})\}$ is used as an estimating function for $\boldsymbol{\psi}$ and the function $D^*_{\boldsymbol{\psi}}\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi})\}$ is used as an estimating function for $\boldsymbol{\xi}$. Theorem 4.2 shows that $\hat{\boldsymbol{\psi}}^{\text{BR}}_n \xrightarrow{p} \boldsymbol{\psi}^*_{\text{BR}}$ and $\hat{\boldsymbol{\xi}}^{\text{BR}}_n \xrightarrow{p} \boldsymbol{\xi}^*_{\text{BR}}$ with $\boldsymbol{\psi}^*_{\text{BR}}=\boldsymbol{\psi}_0$ under $\mathcal{M}(\mathcal{G}_{\boldsymbol{\psi}})$ and $\boldsymbol{\xi}^*_{\text{BR}}=\boldsymbol{\xi}_0$ under $\mathcal{M}(\mathcal{Q}_{\boldsymbol{\xi}})$. This can be understood by noting that the double robustness implies that $P_0\left[D^*_{\boldsymbol{\xi}}\{g(\boldsymbol{\psi}_0),\bar{Q}(\boldsymbol{\xi})\}\right]=\mathbf{0}$ for all $\boldsymbol{\xi}$ under $\mathcal{M}(\mathcal{G}_{\boldsymbol{\psi}})$ and that $P_0\left[D^*_{\boldsymbol{\psi}}\{g(\boldsymbol{\psi}),\bar{Q}(\boldsymbol{\xi}_0)\}\right]=\mathbf{0}$ for all $\boldsymbol{\psi}$ under $\mathcal{M}(\mathcal{Q}_{\boldsymbol{\xi}})$. The bias-reduced doubly robust estimator is now given by $\hat{\mu}_{n,\text{BR}}\equiv\hat{\mu}_{n,\text{DR}}(\hat{g}^{\text{BR}}_n,\hat{\bar{Q}}^{\text{BR}}_n)$, with $\hat{g}^{\text{BR}}_n=\hat{g}(\hat{\boldsymbol{\psi}}^{\text{BR}}_n)$ and $\hat{\bar{Q}}^{\text{BR}}_n=\hat{\bar{Q}}(\hat{\boldsymbol{\xi}}^{\text{BR}}_n)$.

Recall that for the missing data problem discussed in Section 5.2, with working models $\mathcal{G}_{\boldsymbol{\psi}}$ and $\mathcal{Q}_{\boldsymbol{\xi}}$ from Section 5.3.1 but with $\boldsymbol{l}$ and $\boldsymbol{k}$ of the same dimension (i.e., such that $s=r$), estimators for the nuisance parameters $(\hat{\boldsymbol{\psi}}^{\text{BR},T}_n,\hat{\boldsymbol{\xi}}^{\text{BR},T}_n)^T$ are then obtained by solving (5.7) and (5.8),

$$n^{-1}\sum_{i=1}^{n}\left\{1-\frac{R_i}{g(\hat{\boldsymbol{\psi}}^{\text{BR}}_n)(\boldsymbol{X}_i)}\right\}Q'\{\hat{\boldsymbol{\xi}}^{\text{BR},T}_n\boldsymbol{k}(\boldsymbol{X}_i)\}\boldsymbol{k}(\boldsymbol{X}_i) = \mathbf{0}, \tag{5.9}$$

$$n^{-1}\sum_{i=1}^{n}R_i\left\{Y_i-\bar{Q}(\hat{\boldsymbol{\xi}}^{\text{BR}}_n)(\boldsymbol{X}_i)\right\}\frac{G'\{\hat{\boldsymbol{\psi}}^{\text{MLE},T}_n\boldsymbol{l}(\boldsymbol{X}_i)\}}{G^2\{\boldsymbol{\psi}^{\text{MLE},T}_n\boldsymbol{l}(\boldsymbol{X}_i)\}}\boldsymbol{l}(\boldsymbol{X}_i) = \mathbf{0}. \tag{5.10}$$

In Section 4.3, we showed how (5.9) can be solved as an optimization problem via the use of an integrated estimating equation; (5.10) can be solved via weighted least squares based on the complete cases. We refer to Section 4.3 for a detailed discussion on the implications of estimating equations (5.9) and (5.10) on the

behavior of $\hat{\mu}_{n,\text{BR}}$.

A limitation of the bias-reduced estimation procedure is that it is confined to parametric working models of the same dimension. Equal dimensions are indeed required because the gradient of D^* with respect to $\boldsymbol{\xi}$ is used as an estimating function for $\boldsymbol{\psi}$ and vice versa. This can be overcome by minimizing the first-order asymptotic bias in the direction of a single nuisance parameter, for instance in the direction of $\boldsymbol{\psi}$. The effect of minimizing the asymptotic bias in the direction of a single nuisance parameter, rather than both, is best understood from Figure 5.1 (example 1) and Figure 5.2 (example 2). Recall that for example 1, for each individual i,

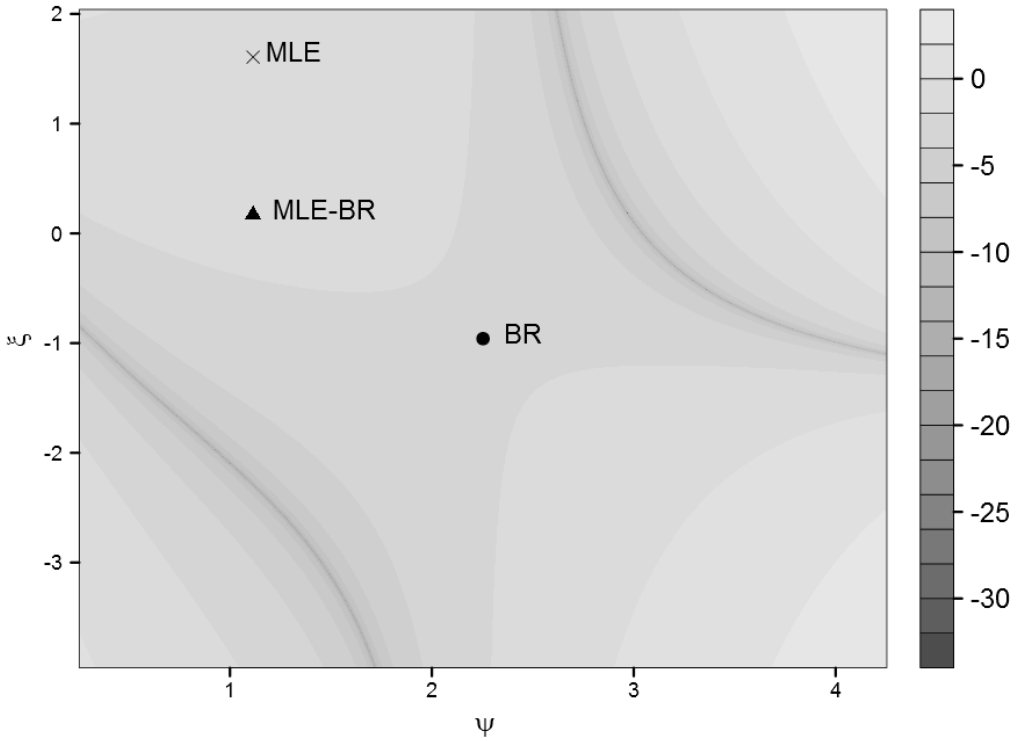


Figure 5.1: *Contourplot of the log of the squared first-order asymptotic bias* $\log\{bias^2(\psi,\xi;1)\}$ *as a function of the nuisance parameters* ψ *and* ξ *for* ***Example 1*** *with* $\times = (1.115, 1.606) \approx (\psi^*_{MLE}, \xi^*_{MLE})$, $\bullet = (2.254, -0.959) \approx (\psi^*_{BR}, \xi^*_{BR})$ *and* $\blacktriangle = (1.115, 0.169) \approx (\psi^*_{MLE}, \xi^*_{MLE\text{-}BR})$.

we have that $X_i \overset{d}{=} N(0,1)$, $R_i|X_i \overset{d}{=} \text{Ber}\{g_0(X_i)\}$ with $g_0(X_i) = \text{expit}(-1+X_i^3)$ and $Y_i|X_i \overset{d}{=} N(X_i^2,1)$, leading to $\mu_0 = 1$. For example 2, for each individual i, we have

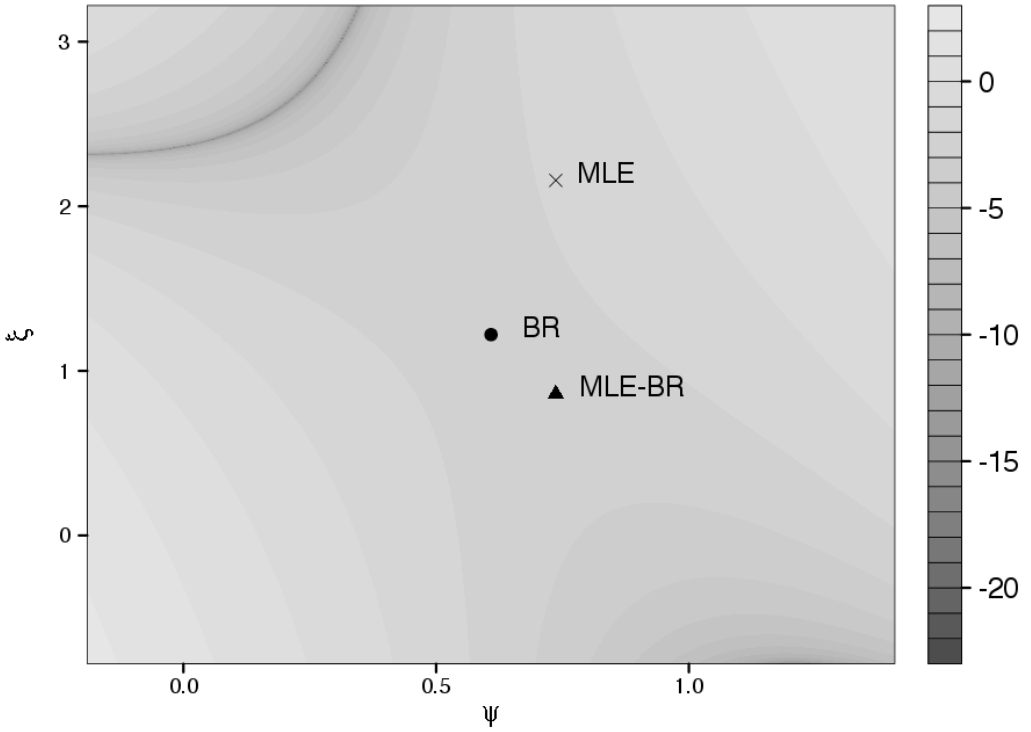


Figure 5.2: *Contourplot of the log of the squared first-order asymptotic bias* $\log\{bias^2(\psi,\xi;2)\}$ *as a function of the nuisance parameters* ψ *and* ξ *for* ***Example 2*** *with* $\times = (0.737, 2.157) \approx (\psi^*_{MLE}, \xi^*_{MLE})$, $\bullet = (0.609, 1.220) \approx (\psi^*_{BR}, \xi^*_{BR})$ *and* $\blacktriangle = (0.737, 0.859) \approx (\psi^*_{MLE}, \xi^*_{MLE\text{-}BR})$.

that $X_i \overset{d}{=} N(1,1)$, $R_i|X_i \overset{d}{=} \text{Ber}\{g_0(X_i)\}$ with $g_0(X_i) = \text{expit}(-1+X_i^2)$ and $Y_i|X_i \overset{d}{=} N(X_i^2,1)$, leading to $\mu_0 = 2$. For both examples, misspecified (one-dimensional) working models are of the form $g(\psi)(X) = \text{expit}(\psi X)$ and $\bar{Q}(\xi)(X) = \xi X$ and we let $n = 10^5$ so that we can ignore sampling variability. Table 5.1 summarizes the results (see also Section 4.3.2). The interpretation for bias reduction in one dimension is then as follows. For a given value $\tilde{\psi}$ of ψ, e.g., the MLE $\hat{\psi}_n^{\text{MLE}}$, and for each value of ξ, we evaluate how the bias of the doubly robust estimator changes as we move away from the chosen value $\tilde{\psi}$. We then choose the outcome regression parameter ξ at which no change is seen. For example 1, with MLE $\hat{\psi}_n^{\text{MLE}} = 1.115$, this leads to $\hat{\xi}_n^{\text{MLE-BR}} = 0.169$, leading to $\hat{\psi}_{n,\text{MLE-BR}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE-BR}}) = 0.518$ and $\text{bias}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE-BR}}; 1) = -0.482$. On Figure 5.1, this point $(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE-BR}}) =$

$(1.115, 0.169)$ is indicated by means of the triangle '▲'. This is the point where no change of bias is seen in the direction of ψ. Note that the bias lies in-between the bias of the MLE and the bias-reduced doubly robust estimator. A qualitatively different result is seen for example 2. With MLE $\hat{\psi}_n^{\text{MLE}} = 0.737$, this leads to $\hat{\xi}_n^{\text{MLE-BR}} = 0.859$, leading to $\hat{\psi}_{n,\text{MLE-BR}} \equiv \hat{\mu}_{n,\text{DR}}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE-BR}}) = 2.302$ and $\text{bias}(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE-BR}}; 2) = 0.302$, which is slightly smaller than that of the bias-reduced doubly robust estimator. This is an artefact of the true underlying data-generating mechanism P_0; there is no theoretical reason for this bias to be smaller (see example 1 where it is higher). On Figure 5.2, this point $(\hat{\psi}_n^{\text{MLE}}, \hat{\xi}_n^{\text{MLE-BR}}) = (0.737, 0.859)$ is also indicated by means of the triangle '▲' and represents the point where no change of bias is seen in the direction of ψ. As before, this strategy does not ensure globally minimal bias. Our proposal ensures instead that small changes in ψ will not induce more bias than what is currently attained.

5

Table 5.1: *Summary results of graphical illustration.*

STRAT ψ	STRAT ξ	$\hat{\psi}_n$	$\hat{\xi}_n$	$\text{bias}(\hat{\psi}_n, \hat{\xi}_n; \mu_0)$	$\hat{\mu}_{n,\text{DR}}(\hat{\psi}_n, \hat{\xi}_n)$
Example 1, $\mu_0 = 1$					
MLE	MLE	1.115	1.606	−0.763	0.237
BR	BR	2.254	−0.959	−0.337	0.663
MLE	BR	1.115	0.160	−0.482	0.518
Example 2, $\mu_0 = 2$					
MLE	MLE	0.737	2.157	0.393	2.393
BR	BR	0.609	1.220	0.316	2.316
MLE	BR	0.737	0.859	0.302	2.302

NOTE: STRAT, estimation strategy; MLE, maximum likelihood estimation; BR, bias-reduced estimation.

5.4 Extending Bias-reduced Doubly Robust Estimation Beyond Parametric Working Models

In this section, we will relax the restriction to parametric working models by making use of data-adaptive learning algorithms to estimate the conditional mean outcome, as in the TMLE procedure. We will however continue to work with a parametric working model for the propensity score because of the concern that a too flexible, data-adaptive model specification may result in near-positivity violations and thereby distort the performance of the doubly robust estimator. With the concern for bias under such parametric working model, we will apply the bias-reduction principle of Section 5.3.2 (see also Chapter 4) in the direction of the nuisance parameters indexing that working model. A side effect of our proposed procedure will be that it no longer constrains the dimensions of both working models to be the same.

5.4.1 Main idea

As previously suggested, the proposed procedure proceeds by postulating a parametric model for the propensity score $g_0(\boldsymbol{X})$, indexed by $\boldsymbol{\psi}$, and obtains an estimator $\hat{\boldsymbol{\psi}}_n$, e.g., via MLE. Next, an estimator for the conditional mean outcome $\bar{Q}_0(\boldsymbol{X})$ is obtained, either by maximum likelihood estimation under a parametric model or by using data-adaptive learning algorithms such as super-learning (van der Laan et al. 2007). With the concern for misspecification of the propensity score model, we next fluctuate this initial estimator of $\bar{Q}_0(\boldsymbol{X})$ through a parametric fluctuation model, indexed by a finite-dimensional parameter $\boldsymbol{\varepsilon}$ of at least the dimension of $\boldsymbol{\psi}$. This model includes covariates that are chosen in such a way that the score of the fluctuation parameter $\boldsymbol{\varepsilon}$ equals the gradient $D^*_{\boldsymbol{\psi}}\{g(\boldsymbol{\psi}), \bar{Q}\}$, so as to minimize the first-order asymptotic bias of the doubly robust estimator in the direction of $\boldsymbol{\psi}$. In Section 5.4.2, we will detail how this can be done. Note that we can thus allow for a propensity score model of arbitrary dimension, at the expense of bias-reduction in only one direction.

The idea of extending an initial fit of the conditional mean outcome is not new: it also underlies the TMLE procedure of van der Laan and Rubin (2006), as

well as other improved doubly robust estimation procedures (Tan 2010; Rotnitzky et al. 2012). Our proposal is nonetheless different in that it explicitly targets bias reduction. In contrast, the TMLE procedure aims at obtaining a doubly robust substitution estimator and the procedures by Tan (2010) and Rotnitzky et al. (2012) target a bounded doubly robust estimator with desirable efficiency properties.

5.4.2 Practical implementation of the procedure

The extension of the bias-reduced doubly robust estimation procedure can be implemented using the following four steps.

Step 1: Estimator $\hat{g}_n^{\text{MLE}}$ for the propensity score $g_0(\boldsymbol{X})$. Postulate a parametric working model for $g_0(\boldsymbol{X})$: $\mathcal{G}_{\boldsymbol{\psi}} = \{g(\boldsymbol{\psi})(\boldsymbol{X}) | \boldsymbol{\psi} \in \mathbb{R}^s\}$ where $g(\boldsymbol{\psi})(\boldsymbol{X}) = G\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\}$ and G is an appropriate inverse link function and $\boldsymbol{l} = (1, l_1, \ldots, l_{s-1})$, e.g., the logistic regression model $G(\cdot) = \text{expit}(\cdot)$. Obtain the MLE $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ solving

$$P_n \left\{ D_g^{\text{MLE}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}) \right\} = \mathbf{0},$$

with $D_g^{\text{MLE}}(\boldsymbol{\psi})$ given by (5.5). Let $\hat{g}_n^{\text{MLE}} = g(\hat{\boldsymbol{\psi}}_n^{\text{MLE}})$.

Step 2: Initial estimator $\hat{\bar{Q}}_n^0$ for the conditional mean outcome $\bar{Q}_0$. The second step of the procedure is to obtain an initial estimator for $\bar{Q}_0$. We describe two possibilities: (1) a parametric model and (2) a super-learner.

1. *Parametric Working Model.* The first option is to postulate a parametric working model for $\bar{Q}_0(\boldsymbol{X})$: $\mathcal{Q}_{\boldsymbol{\xi}} = \{\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) | \boldsymbol{\xi} \in \mathbb{R}^r\}$ with $\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) = Q\{\boldsymbol{\xi}^T \boldsymbol{k}(\boldsymbol{X})\}$, Q an appropriate inverse link function and $\boldsymbol{k} = (1, k_1, \ldots, k_{r-1})$; e.g., a linear regression model with Q the identity link. Obtain the MLE $\hat{\boldsymbol{\xi}}_n^{\text{MLE}}$ solving
$$P_n \left\{ D_{\bar{Q}}^{\text{MLE}}(\hat{\boldsymbol{\xi}}_n^{\text{MLE}}) \right\} = \mathbf{0},$$
with $D_{\bar{Q}}^{\text{MLE}}(\boldsymbol{\xi})$ given by (5.6). Let $\hat{\bar{Q}}_n^{\text{MLE}} = \bar{Q}(\hat{\boldsymbol{\xi}}_n^{\text{MLE}})$.

2. *Super-Learner.* Another option is to obtain an initial estimator based on data-adaptive learning algorithms, such as super-learning (van der Laan

et al. 2007). The super-learner is a machine learning algorithm which starts from a library of estimators $\{\hat{\bar{Q}}_j | j = 1,\ldots,J\}$, which may consist of both parametric and nonparametric estimators. It then considers the family of all weighted averages of these estimators: $\hat{\bar{Q}}_{\boldsymbol{\omega}} = \sum_{j=1}^{J} \omega_j \hat{\bar{Q}}_j$, $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_J)^T$, $\sum_{j=1}^{J} \omega_j = 1$ and $\omega_j \geq 0$ for $j = 1,\ldots,J$. Next, the optimal weight vector $\hat{\boldsymbol{\omega}}_n$ is defined to be the choice of $\boldsymbol{\omega}$ that minimizes the cross-validated risk with respect to some loss-function $\mathcal{L}_{\text{SL}} : (\boldsymbol{O}, \bar{Q}) \to \mathcal{L}_{\text{SL}}(\bar{Q})(\boldsymbol{O}) \in \mathbb{R}$ which satisfies $\bar{Q}_0 = \arg\min_{\bar{Q}} E\{\mathcal{L}_{\text{SL}}(\bar{Q})\}$, e.g., the squared error loss-function $\mathcal{L}_2(\bar{Q})(\boldsymbol{O}) = R\{Y - \bar{Q}(\boldsymbol{X})\}^2$. The super-learner of $\bar{Q}_0$ is defined as the estimator $\hat{\bar{Q}}_n^{\text{SL}} = \hat{\bar{Q}}_{\hat{\boldsymbol{\omega}}_n}$.

For further reference, we let the initial estimator $\hat{\bar{Q}}_n^0$ denote either $\hat{\bar{Q}}_n^{\text{MLE}}$ or $\hat{\bar{Q}}_n^{\text{SL}}$.

Step 3: Fluctuation $\hat{\bar{Q}}_n^{(c)}$ of the initial estimator $\hat{\bar{Q}}_n^0$. To construct an appropriate fluctuation model, we need to choose an appropriate loss-function. We consider the quasi-log-likelihood loss-function with corresponding logistic fluctuation model. The favoured choice of the quasi-log-likelihood loss-function with logistic fluctuation model over the least-squares loss-function with linear fluctuation model is well-studied in Gruber and van der Laan (2010). It is found to be favourable because it inherits predictions within the admissible range for the outcome. For this purpose, suppose Y is known to fall in the interval $[a,b]$. To be able to use the quasi-log-likelihood loss-function and the logistic fluctuation model, we need the outcome to fall between zero and one. This can be easily accomplished by considering the linearly transformed outcome

$$\widetilde{Y} = \frac{Y-a}{b-a}. \tag{5.11}$$

The procedure we describe below is based on the transformed outcome $\widetilde{Y}$ and results in an estimator $\hat{\tilde{\mu}}_n$ of $\tilde{\mu}_0 = E(\widetilde{Y})$. Because $E(Y) = (b-a)E(\widetilde{Y}) + a$, the final estimator is given by $\hat{\mu}_n = (b-a)\hat{\tilde{\mu}}_n + a$. For notational convenience, without loss of generality, we will assume that $a = 0$ and $b = 1$ so that $Y = \widetilde{Y} \in [0,1]$ and we can drop the $\sim$-notation.

Having thus obtained an estimator for the propensity score, such as $\hat{g}_n^{\text{MLE}}$, and an

initial estimator $\hat{\bar{Q}}_n^0$ for the conditional mean outcome taking values in the interval $[0,1]$, we can now fluctuate $\hat{\bar{Q}}_n^0$ such that the squared first-order asymptotic bias of the doubly robust estimator is minimized in the direction of the finite-dimensional parameter $\boldsymbol{\psi}$. Below, we consider three fluctuation models that will accomplish this goal, with the second also guaranteeing the final estimator to be a substitution estimator, like the TMLE procedure.

1. *Fluctuation model 1.* Consider the fluctuation model $\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(1)}) : \boldsymbol{\varepsilon}^{(1)} \in \mathbb{R}^s\right\}$ through the initial estimator ($\hat{\bar{Q}}_n^0(\mathbf{0}) = \hat{\bar{Q}}_n^0$):

$$\operatorname{logit}\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(1)})(\boldsymbol{X}) = \operatorname{logit}\hat{\bar{Q}}_n^0(\boldsymbol{X}) + \boldsymbol{\varepsilon}^{(1),T}\boldsymbol{H}^{(1)}(\hat{g}_n^{\mathrm{MLE}})(\boldsymbol{X}), \tag{5.12}$$

with $\boldsymbol{H}^{(1)}(\hat{g}_n^{\mathrm{MLE}})(\boldsymbol{X}) = [G'\{\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE},T}\boldsymbol{l}(\boldsymbol{X})\}/G^2\{\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE},T}\boldsymbol{l}(\boldsymbol{X})\}]\boldsymbol{l}(\boldsymbol{X})$ and with $\operatorname{logit}(x) = \log\{x/(1-x)\}$. Then define $\hat{\boldsymbol{\varepsilon}}_n^{(1)}$ such that

$$\hat{\boldsymbol{\varepsilon}}_n^{(1)} = \underset{\boldsymbol{\varepsilon}^{(1)}}{\arg\min}\, P_n\left[\mathcal{L}^{(1)}\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(1)})\right\}\right]$$

where $\mathcal{L}^{(1)}$ is the quasi-log-likelihood loss-function;

$$\mathcal{L}^{(1)}(\bar{Q})(\boldsymbol{O}) = -R[Y\log\{\bar{Q}(\boldsymbol{X})\} + (1-Y)\log\{1-\bar{Q}(\boldsymbol{X})\}]. \tag{5.13}$$

It is easily verified that $\hat{\boldsymbol{\varepsilon}}_n^{(1)}$ solves the estimating equation

$$\sum_{i=1}^{n} R_i\left\{Y_i - \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(1)})(\boldsymbol{X}_i)\right\}\boldsymbol{H}^{(1)}(\hat{g}_n^{\mathrm{MLE}})(\boldsymbol{X}_i) = \mathbf{0}, \tag{5.14}$$

which can be solved via standard logistic regression of the observed outcomes on the covariates $\boldsymbol{H}^{(1)}$ using as offset $\operatorname{logit}\hat{\bar{Q}}_n^0$. The score equation (5.14) is like the estimation equation (5.10) and therefore guarantees bias-reduction (in the direction of $\boldsymbol{\psi}$). Define the updated estimator $\hat{\bar{Q}}_n^{(1)} = \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(1)})$. Because the quasi-log-likelihood loss-function is a valid loss-function for the conditional mean outcome $\bar{Q}_0$ in the sense that $\bar{Q}_0 = \arg\min_{\bar{Q}} E\{\mathcal{L}^{(1)}(\bar{Q})\}$ (see Gruber and van der Laan (2010), Lemma 1), $\hat{\bar{Q}}_n^{(1)}$ is an improved fit of $\bar{Q}_0$ as compared to $\hat{\bar{Q}}_n^0$ with respect to the loss-function (5.13).

2. *Fluctuation model 2.* We extend (5.12) so that the final doubly robust estimator of the target parameter will also be a substitution estimator (also known as a regression doubly robust estimator (Robins et al. 2007)). For this purpose, define the fluctuation model $\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(2)}) : \boldsymbol{\varepsilon}^{(2)} \in \mathbb{R}^{s+1}\right\}$ through the initial estimator:

$$\begin{aligned}&\operatorname{logit}\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(2)})(\boldsymbol{X})\\&= \operatorname{logit}\hat{\bar{Q}}_n^0(\boldsymbol{X}) + \boldsymbol{\varepsilon}^{(2,1),T}\boldsymbol{H}^{(1)}(\hat{g}_n^{\text{MLE}})(\boldsymbol{X}) + \varepsilon^{(2,2)}H^{(2)}(\hat{g}_n^{\text{MLE}})(\boldsymbol{X}),\end{aligned} \tag{5.15}$$

$\boldsymbol{\varepsilon}^{(2)} = (\boldsymbol{\varepsilon}^{(2,1),T}, \varepsilon^{(2,2)})^T$ and $H^{(2)}(\hat{g}_n^{\text{MLE}}) = 1/\hat{g}_n^{\text{MLE}}$. Then define $\hat{\boldsymbol{\varepsilon}}_n^{(2)}$ such that

$$\hat{\boldsymbol{\varepsilon}}_n^{(2)} = \underset{\boldsymbol{\varepsilon}^{(2)}}{\arg\min}\, P_n\left[\mathcal{L}^{(1)}\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(2)})\right\}\right]$$

where $\mathcal{L}^{(1)}$ is the quasi-log-likelihood loss-function (5.13). It follows that $\hat{\boldsymbol{\varepsilon}}_n^{(2)}$ solves the estimating equation

$$\begin{aligned}&\sum_{i=1}^n R_i\left\{Y_i - \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(2)})(\boldsymbol{X}_i)\right\}\\&\qquad\times\left(\boldsymbol{H}^{(1),T}(\hat{g}_n^{\text{MLE}})(\boldsymbol{X}_i), H^{(2)}(\hat{g}_n^{\text{MLE}})(\boldsymbol{X}_i)\right)^T = \boldsymbol{0},\end{aligned} \tag{5.16}$$

which can be easily solved via standard logistic regression of the observed outcomes on the covariates $\boldsymbol{H}^{(1)}$ and $H^{(2)}$ using as offset $\operatorname{logit}\hat{\bar{Q}}_n^0$. By adding $H^{(2)}$ to the fluctuation model, it follows from the second component of (5.16) that the final estimator will be a substitution estimator (as in the TMLE procedure). This is because (5.16) implies that $\sum_{i=1}^n R_i/\hat{g}_n^{\text{MLE}}\left\{Y_i - \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(2)})(\boldsymbol{X}_i)\right\} = 0$ so that $\hat{\mu}_{n,\text{DR}}\{\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(2)})\} = P_n\{\hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(2)})\}$. Define the updated estimator $\hat{\bar{Q}}_n^{(2)} = \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(2)})$.

3. *Fluctuation model 3.* There is evidence in the literature that the use of a weighted loss-function in the construction of a fluctuation model can improve the stability of the doubly robust estimator of interest; see for example Robins et al. (2007), and see Díaz and Rosenblum (2014) for a comparison of an unweighted versus a weighted loss-function in the context of TMLE.

We therefore propose the following alternative fluctuation model to (5.12): $\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(3)}) : \boldsymbol{\varepsilon}^{(3)} \in \mathbb{R}^s\right\}$ through the initial estimator:

$$\operatorname{logit}\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(3)})(\boldsymbol{X}) = \operatorname{logit}\hat{\bar{Q}}_n^0(\boldsymbol{X}) + \boldsymbol{\varepsilon}^{(3),T}\boldsymbol{l}(\boldsymbol{X}). \tag{5.17}$$

Then define $\hat{\boldsymbol{\varepsilon}}_n^{(3)}$ such that

$$\hat{\boldsymbol{\varepsilon}}_n^{(3)} = \underset{\boldsymbol{\varepsilon}^{(3)}}{\arg\min} P_n\left[\frac{G'\{\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE},T}\boldsymbol{l}(\boldsymbol{X})\}}{G^2\{\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE},T}\boldsymbol{l}(\boldsymbol{X})\}}\mathcal{L}^{(1)}\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}^{(3)})\right\}\right],$$

with $\mathcal{L}^{(1)}$ the quasi-log-likelihood loss-function (5.13); thus $\hat{\boldsymbol{\varepsilon}}_n^{(3)}$ solves the estimating equations

$$0 = \sum_{i=1}^n R_i\left\{Y_i - \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(2)})(\boldsymbol{X}_i)\right\}\frac{G'\{\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE},T}\boldsymbol{l}(\boldsymbol{X}_i)\}}{G^2\{\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE},T}\boldsymbol{l}(\boldsymbol{X}_i)\}}\boldsymbol{l}(\boldsymbol{X}_i), \tag{5.18}$$

which can be easily solved via standard weighted logistic regression of the observed outcomes on the covariates $\boldsymbol{l}$ using as offset $\operatorname{logit}\hat{\bar{Q}}_n^0$. The equation (5.18) is again like the estimating equation (5.10) and therefore guarantees bias-reduction (in the direction of $\boldsymbol{\psi}$). Then define the updated estimator $\hat{\bar{Q}}_n^{(3)} = \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(3)})$.

We thus obtain three different updated estimators $\hat{\bar{Q}}_n^{(c)}$, where c will denote either 1, 2 or 3. These updated estimators all share the same property of reducing bias in the direction of $\boldsymbol{\psi}$. As noted, this is because each of the scores for the fluctuation parameters $\boldsymbol{\varepsilon}^{(c)}$ $(c = 1,2,3)$ satisfy the property that

$$P_n\left\{D^*_{\boldsymbol{\psi}}(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)})\right\} = \mathbf{0}.$$

From Theorem 4.1, it thus follows that the fluctuation parameter $\boldsymbol{\varepsilon}^{(c)}$ is estimated in a way that the squared first-order asymptotic bias of the doubly robust estimator as a function of $\boldsymbol{\psi}$ is locally minimal in the probability limit of $\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE}}$.

Step 4: Estimating the Target Parameter $\hat{\mu}_n^{(c)}$. Given the estimators $\hat{g}_n^{\mathrm{MLE}}$ and $\hat{\bar{Q}}_n^{(c)}$, we obtain the doubly robust estimator $\hat{\mu}_n^{(c)} \equiv \hat{\mu}_{n,\mathrm{DR}}(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)})$. Note that by

construction (implied by the estimating equation (5.16) as a consequence of adding $H^{(2)}$ to the fluctuation model), $\hat{\mu}_n^{(2)} = \hat{\mu}_{n,\text{DR}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(2)}) = P_n(\hat{\bar{Q}}_n^{(2)})$.

Remark 5.1. *If in the construction of the propensity score model, we take $G(x) = expit(x)$, then $G'(x) = G(x)\{1-G(x)\}$. For every $c = 1,2,3$, it then follows from the estimating equations for the fluctuation parameter $\boldsymbol{\varepsilon}^{(c)}$ that*

$$\sum_{i=1}^{n} \frac{R_i}{\hat{g}_n^{MLE}(\boldsymbol{X}_i)} \left\{ Y_i - \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i) \right\} = \sum_{i=1}^{n} R_i \left\{ Y_i - \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i) \right\}. \tag{5.19}$$

This implies that the doubly robust estimator can be written as

$$\hat{\mu}_n^{(c)} = n^{-1} \sum_{i=1}^{n} \left\{ R_i Y_i + (1-R_i) \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i) \right\}, \tag{5.20}$$

which averages the observed outcome for the responders and a predicted outcome for the non-responders, making the doubly robust estimator equal to a special type of imputation estimator. This is desirable as this ensures that the doubly robust estimator is sample bounded in the sense that $\hat{\mu}_n^{(c)}$ lies in the observed data range (Robins et al. 2007; Tan 2010).

5

5.4.3 Inference

In Appendix 5.A, we present an asymptotic linearity theorem with corresponding influence function for the doubly robust estimators $\hat{\mu}_n^{(c)}$ $(c = 1,2,3)$ under the assumption of a correctly specified propensity score working model $\mathcal{M}(\mathcal{G})$ but a potentially misspecified working model for the conditional mean outcome $\mathcal{M}(\mathcal{Q})$. This asymptotic linearity of $\hat{\mu}_n^{(c)}$ and the corresponding influence function

$$D^*(g_0, \bar{Q}_{(c)}^*; \mu_0),$$

with $\bar{Q}_{(c)}^*$ the probability limit of $\hat{\bar{Q}}_n^{(c)}$, provides us with a strategy for inference about the unknown μ_0 under a correctly specified propensity score model. Note that when $\bar{Q}_{(c)}^* = \bar{Q}_0$ and thus the outcome model is correctly specified, the influence function equals $D^*(g_0, \bar{Q}_0; \mu_0)$, which is the efficient influence function. If $\bar{Q}_{(c)}^* \neq \bar{Q}_0$ and

thus the outcome model is misspecified, the influence function of the doubly robust estimator equals $D^*(g_0, \bar{Q}^*_{(c)}; \mu_0)$ and does not equal the efficient influence function (indeed, $\hat{\mu}_n^{(c)}$ is only locally efficient, efficiency is only attained locally when $\bar{Q}^*_{(c)} = \bar{Q}_0$). This will not lead to conservative inference by not acknowledging the uncertainty of the estimator for the propensity score in the influence function calculation. This is because bias-reduced estimation is used in the direction of $\boldsymbol{\psi}$ implying that

$$P_0\left\{D^*_{\boldsymbol{\psi}}(g_0, \bar{Q}^*_{(c)})\right\} = \mathbf{0}.$$

This first-order ancillarity property with respect to $\boldsymbol{\psi}$ ensures that the resulting doubly robust estimator will be insensitive to local changes (of the order one over root-n) in the nuisance parameter $\boldsymbol{\psi}$, even under misspecification of the outcome model. Consequently, no correction for the estimation of the propensity score is needed under inconsistency of $\hat{\bar{Q}}_n^{(c)}$, an artefact of the bias-reduced estimation strategy.

5

Standard errors. When the propensity score model is correctly specified, a standard error for $\hat{\mu}_n^{(c)}$ can be easily calculated as the square root of the sample variance of the estimated influence function divided by n:

$$\widehat{\mathrm{SE}}(\hat{\mu}_n^{(c)}) = \sqrt{\frac{\widehat{\mathrm{var}}\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \hat{\mu}_n^{(c)})\right\}}{n}} \tag{5.21}$$

with

$$\begin{aligned}
&\widehat{\mathrm{var}}\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \hat{\mu}_n^{(c)})\right\} \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \hat{\mu}_n^{(c)})(\boldsymbol{O}_i)\right\}^2 \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i}{\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X}_i)}\left\{Y_i - \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i)\right\} + \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i) - \hat{\mu}_n^{(c)}\right]^2.
\end{aligned}$$

An alternative for standard error calculation, which does not demand the assumption that the propensity score model is correctly specified, is to use the nonparametric bootstrap. However, there is no theory supporting that the nonparametric bootstrap

would produce valid results when the estimators rely on data-adaptive estimation such as super-learning (van der Laan 2014).

Confidence intervals and p-values. Given the estimator $\widehat{\mathrm{SE}}(\hat{\mu}_n^{(c)})$ for the standard error of $\hat{\mu}_n^{(c)}$, a confidence interval and p-value can be calculated based on the asymptotic normality of the estimator. A $(1-\alpha)100\%$ CI is given by

$$\left[\hat{\mu}_n^{(c)} \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\mu}_n^{(c)})\right]$$

where $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1-\alpha/2$ and $\Phi(\cdot)$ the cumulative distribution function of a standard normal random variable. A p-value for the hypothesis test $H_0 : \mu = \tilde{\mu}$ versus $H_a : \mu_0 \neq \tilde{\mu}$ can be calculated as

$$p = 2\left\{1 - \Phi\left(\left|\frac{\hat{\mu}_n^{(c)} - \tilde{\mu}}{\widehat{\mathrm{SE}}(\hat{\mu}_n^{(c)})}\right|\right)\right\}.$$

Inference on the original scale. With $\widetilde{Y} = (Y-a)/(b-a)$, the transformed data which falls within the interval $[0,1]$, previously, we let $\hat{\tilde{\mu}}_n^{(c)}$ denote the final estimator for $\tilde{\mu}_0 = E(\widetilde{Y})$. Because $E(Y) = (b-a)E(\widetilde{Y}) + a$, the final estimator for $\mu_0 = E(Y)$ can be obtained as $\hat{\mu}_n^{(c)} = (b-a)\hat{\tilde{\mu}}_n^{(c)} + a$. When $n^{1/2}(\hat{\tilde{\mu}}_n^{(c)} - \tilde{\mu}_0)$ converges in distribution to $N(0,\tilde{\sigma}^2)$, then $n^{1/2}(\hat{\mu}_n^{(c)} - \mu_0)$ will converge in distribution to $N(0,\sigma^2)$ with $\sigma^2 = (b-a)^2\tilde{\sigma}^2$. Consequently, a standard error for $\hat{\mu}_n^{(c)}$ on the original scale of the outcome, can be obtained as $\widehat{\mathrm{SE}}(\hat{\mu}_n^{(c)}) = (b-a)\widehat{\mathrm{SE}}(\hat{\tilde{\mu}}_n^{(c)})$.

5.5 Simulation Studies

We carried out different simulation studies to compare the performance of the new bias-reduced doubly robust estimators with several alternatives for the estimation of a mean outcome in the presence of incomplete data.

5.5.1 Estimators

All estimators are based on a parametric working model $\mathcal{G}_{\boldsymbol{\psi}}$ for the propensity score $g_0(\boldsymbol{X})$. Let $\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE}}$ denote the MLE and let $\hat{g}_n^{\mathrm{MLE}} = g(\hat{\boldsymbol{\psi}}_n^{\mathrm{MLE}})$. For the conditional

mean outcome $\bar{Q}_0(\boldsymbol{X})$, we both consider estimators based on a parametric working model and based on data-adaptive learning algorithms. First, consider a parametric model $\mathcal{Q}_{\boldsymbol{\xi}}$ with $\hat{\boldsymbol{\xi}}_n^{\text{MLE}}$ the MLE and $\hat{\bar{Q}}_n^{\text{MLE}} = \bar{Q}(\hat{\boldsymbol{\xi}}_n^{\text{MLE}})$. Given the parametric working models $\mathcal{G}_{\boldsymbol{\psi}}$ and $\mathcal{Q}_{\boldsymbol{\xi}}$, let $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\xi}}_n^{\text{BR}}$ denote the nuisance parameter estimators for $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ obtained via the bias-reduced estimation principle with $\hat{g}_n^{\text{BR}} = g(\hat{\boldsymbol{\psi}}_n^{\text{BR}})$ and $\hat{\bar{Q}}_n^{\text{BR}} = \bar{Q}(\hat{\boldsymbol{\xi}}_n^{\text{BR}})$. Second, we consider the super-learner $\hat{\bar{Q}}_n^{\text{SL}}$ based on a library consisting of generalized additive and linear models, random forests and adaptive polynomial splines. This can be fitted using the `SuperLearner` R package (Polley and van der Laan 2014). Let $\hat{\bar{Q}}_n^{\text{MLE},(c)} = \hat{\bar{Q}}_n^{\text{MLE}}(\hat{\boldsymbol{\varepsilon}}_n^{(c)})$ denote the updated estimators for the conditional mean outcome based on the initial estimator $\hat{\bar{Q}}_n^{\text{MLE}}$, $c = 1,2,3$ and let $\hat{\bar{Q}}_n^{\text{SL},(c)} = \hat{\bar{Q}}_n^{\text{SL}}(\hat{\boldsymbol{\varepsilon}}_n^{(c)})$ denote the updated estimators for the conditional mean outcome based on the initial estimator $\hat{\bar{Q}}_n^{\text{SL}}$, $c = 1,2,3$. The estimators under consideration are then the doubly robust estimator based on the MLE for the parametric working models $\hat{\mu}_{n,\text{MLE}} = \hat{\mu}_{n,\text{DR}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{\text{MLE}})$, the bias-reduced doubly robust estimator $\hat{\mu}_{n,\text{BR}} = \hat{\mu}_{n,\text{DR}}(\hat{g}_n^{\text{BR}}, \hat{\bar{Q}}_n^{\text{BR}})$, the substitution estimator $\hat{\mu}_{n,\text{SL}} = P_n(\hat{\bar{Q}}_n^{\text{SL}})$ based on standardizing super-learning predictions (which is not doubly robust), the new estimators based on both a parametric working model for the propensity score and the conditional mean outcome $\hat{\mu}_{n,\text{MLE}}^{(c)} = \hat{\mu}_{n,\text{DR}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{\text{MLE},(c)})$ and the new estimators based on a parametric model for the propensity score model but a super-learner for the conditional mean outcome $\hat{\mu}_{n,\text{SL}}^{(c)} = \hat{\mu}_{n,\text{DR}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{\text{SL},(c)})$, both for all three fluctuation models $c = 1,2,3$. Finally, we consider two versions of TMLE, described in van der Laan and Rubin (2006) and Gruber and van der Laan (2010). Both are based on the parametric propensity score model $\mathcal{G}_{\boldsymbol{\psi}}$ fitted via MLE but the first is based on $\hat{\bar{Q}}_n^{\text{MLE}}$ and the second is based on $\hat{\bar{Q}}_n^{\text{SL}}$, leading to the TMLEs $\hat{\mu}_{n,\text{TMLE}}$ and $\hat{\mu}_{n,\text{SL-TMLE}}$, respectively.

There are many other alternative estimation strategies for these nuisance working models considered in the literature (see Section 3.5 for an overview). For a simulation-based comparison of many of these alternatives, we refer to Tan (2010), Porter et al. (2011) and Vermeulen and Vansteelandt (2015a), see also Section 4.4.

Remark 5.2. *In the two simulation settings of the subsequent sections, the outcome Y is not a priori bounded. For those estimators involving a logistic fluctuation model for the initial estimator of the conditional mean outcome, we follow the default setting in the* `tmle` *R package (Gruber and van der Laan 2014) by taking the bounds to be the observed range of Y but widened by 10% of both the minimum and maximum value. This differs slightly from the approach in Gruber and van der Laan (2010). Specifically, we let $a = \min_{i=1}^{n} Y_i - 0.1|\min_{i=1}^{n} Y_i|$ and $b = \max_{i=1}^{n} Y_i + 0.1|\max_{i=1}^{n} Y_i|$ and $\widetilde{Y} = (Y-a)/(b-a)$. Next, since the initial estimator needs to be represented as a logistic function of its logit transformation, the initial estimator needs to be bounded away from 0 and 1 because $\text{logit}(x)$ is not defined for $x = 0$ or 1. Therefore, the initial estimator is truncated at $(\zeta, 1-\zeta)$ for some small $\zeta > 0$. In the simulations, we take $\zeta = 0.005$ as in Gruber and van der Laan (2010).*

For each of the scenarios considered below, we perform 1000 Monte Carlo runs at sample sizes of $n = 200$ and 1000. For each estimator, we calculated the Monte Carlo bias (BIAS), the root mean square error (RMSE) and the Monte Carlo standard deviation (MCSD). For the doubly robust estimators (all but $\hat{\mu}_n^{\text{SL}}$), we also show the average sandwich standard error (ASSE) and the Monte Carlo coverage of the corresponding 95% Wald confidence intervals (COV) where standard errors are calculated via formula (5.21).

5.5.2 Scenario 1: one-covariate setting

Data-generating mechanism

This simulation scenario considers the simple data-generating mechanism of Section 4.4.1 where for each i ($i = 1, \ldots, n$),

$$
\begin{aligned}
&X_i \overset{d}{=} N(0,1),\\
&R_i|X_i \overset{d}{=} \text{Ber}\{g_0(X_i)\} \text{ and}\\
&Y_i|X_i \overset{d}{=} N\{\bar{Q}_0(X_i), 1\}.
\end{aligned}
$$

For each setting, the following parametric working models are used: $g(\boldsymbol{\psi})(X) = \text{expit}(\psi_1 + \psi_2 X)$ and $\bar{Q}(\boldsymbol{\xi})(X) = \xi_1 + \xi_2 X$. Simulation experiments with correctly specified parametric working models used $\bar{Q}_0(X) = 1 + X$ and $g_0(X) = \text{expit}(\gamma X)$ for $\gamma = 1, 2$ (see Figure 4.3). To allow for misspecification in the outcome model, we additionally generated data using $\bar{Q}_0(X) = X^2$ and $g_0(X) = \text{expit}(\gamma X)$ for $\gamma = 1, 2$. To allow for misspecification of the propensity score model, we generated data using $\bar{Q}_0(X) = 1 + X$ and $g_0(X) = \text{expit}(-4 + 1.5|X|^{0.5} + 0.75X + 0.5|X|^{1.5})$ (see Figure 4.4), as in Vansteelandt et al. (2012). Finally, we also generated data with $\bar{Q}_0(X) = X^2$ and $g_0(X) = \text{expit}(-4 + 1.5|X|^{0.5} + 0.75X + 0.5|X|^{1.5})$ to allow for misspecification of both models. In each of the settings, the target parameter $E(Y) = \mu_0$ equals one. Table 5.2 shows for each underlying propensity score, the probability $P(R = 0)$, that is, the marginal probability of the outcome Y being missing.

Table 5.2: *Marginal probability of the outcome being missing.*

PROPENSITY SCORE	$P(R=0)$
$g_0(X) = \text{expit}(X)$	0.50
$g_0(X) = \text{expit}(2X)$	0.52
$g_0(X) = \text{expit}(-4 + 1.5\|X\|^{0.5} + 0.75X + 0.5\|X\|^{1.5})$	0.86

Results for the Scenario 1 are given in Table 5.3 ($n = 200$) and Table 5.4 ($n = 1000$).

Results

When both working models are correctly specified and weights are not highly variable ($\gamma = 1$), all estimators tend to perform similarly in terms of bias and precision. However, when weights become highly variable ($\gamma = 2$), estimators $\hat{\mu}^{(1)}_{n,\text{MLE}}$, $\hat{\mu}^{(2)}_{n,\text{MLE}}$, $\hat{\mu}^{(1)}_{n,\text{SL}}$ and $\hat{\mu}^{(2)}_{n,\text{SL}}$ tend to show some finite-sample bias ($n = 200$), which is resolved when the sample size is increased ($n = 1000$). With highly variable weights, these estimators are also relatively less efficient. When the outcome model is misspecified but the working model for the propensity is correct, we observe adequate performance for all estimators, but smaller bias and larger

Table 5.3: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 1,* $n = 200$.

EST	BIAS	RMSE	MCSD	ASSE	COV	BIAS	RMSE	MCSD	ASSE	COV
				$n = 200$						
	OR correct, PS correct ($\gamma = 1$)					OR incorrect, PS correct ($\gamma = 1$)				
$\hat{\mu}_{n,\text{MLE}}$	−0.001	0.13	0.13	0.13	0.94	−0.018	0.33	0.33	0.29	0.89
$\hat{\mu}_{n,\text{BR}}$	−0.001	0.13	0.13	0.13	0.94	−0.030	0.21	0.20	0.17	0.88
$\hat{\mu}_{n,\text{SL}}$	0.007	0.13	0.13	–	–	−0.085	0.19	0.17	–	–
$\hat{\mu}_{n,\text{TMLE}}$	−0.000	0.13	0.13	0.13	0.94	−0.036	0.27	0.26	0.24	0.89
$\hat{\mu}_{n,\text{SL-TMLE}}$	0.001	0.13	0.13	0.13	0.94	−0.029	0.17	0.17	0.14	0.89
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	0.005	0.14	0.14	0.13	0.93	0.028	0.19	0.19	0.16	0.92
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	0.010	0.14	0.14	0.13	0.93	0.037	0.21	0.21	0.18	0.91
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	−0.000	0.13	0.13	0.13	0.94	−0.055	0.22	0.22	0.18	0.85
$\hat{\mu}_{n,\text{SL}^{(1)}}$	0.004	0.14	0.14	0.13	0.93	−0.000	0.16	0.16	0.14	0.92
$\hat{\mu}_{n,\text{SL}^{(2)}}$	0.010	0.14	0.14	0.13	0.92	0.011	0.17	0.17	0.14	0.92
$\hat{\mu}_{n,\text{SL}^{(3)}}$	0.001	0.13	0.13	0.13	0.94	−0.028	0.17	0.17	0.14	0.89
	OR correct, PS correct ($\gamma = 2$)					OR incorrect, PS correct ($\gamma = 2$)				
$\hat{\mu}_{n,\text{MLE}}$	−0.002	0.20	0.20	0.17	0.94	−0.060	1.19	1.19	0.58	0.68
$\hat{\mu}_{n,\text{BR}}$	−0.001	0.19	0.19	0.15	0.88	−0.110	0.26	0.24	0.17	0.77
$\hat{\mu}_{n,\text{SL}}$	0.018	0.15	0.15	–	–	−0.249	0.33	0.22	–	–
$\hat{\mu}_{n,\text{TMLE}}$	0.017	0.19	0.19	0.15	0.86	0.013	0.33	0.33	0.24	0.82
$\hat{\mu}_{n,\text{SL-TMLE}}$	0.018	0.19	0.19	0.14	0.85	−0.020	0.23	0.23	0.16	0.82
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	0.097	0.30	0.28	0.14	0.70	0.304	0.53	0.43	0.22	0.65
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	0.110	0.31	0.29	0.14	0.67	0.282	0.51	0.43	0.21	0.64
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	0.003	0.17	0.17	0.14	0.88	−0.187	0.38	0.33	0.19	0.61
$\hat{\mu}_{n,\text{SL}^{(1)}}$	0.090	0.28	0.27	0.14	0.70	0.128	0.34	0.32	0.17	0.72
$\hat{\mu}_{n,\text{SL}^{(2)}}$	0.111	0.31	0.29	0.14	0.66	0.100	0.35	0.34	0.17	0.69
$\hat{\mu}_{n,\text{SL}^{(3)}}$	0.010	0.18	0.18	0.14	0.86	−0.093	0.25	0.23	0.15	0.75
	OR correct, PS incorrect					OR incorrect, PS incorrect				
$\hat{\mu}_{n,\text{MLE}}$	−0.008	1.04	1.04	0.76	0.96	5.496	11.46	10.06	4.78	0.90
$\hat{\mu}_{n,\text{BR}}$	−0.003	0.29	0.29	0.22	0.85	1.030	1.23	0.68	0.43	0.34
$\hat{\mu}_{n,\text{SL}}$	0.030	0.29	0.29	–	–	0.155	0.42	0.39	–	–
$\hat{\mu}_{n,\text{TMLE}}$	0.010	0.30	0.30	0.49	0.94	1.199	1.34	0.60	0.94	0.79
$\hat{\mu}_{n,\text{SL-TMLE}}$	−0.002	0.29	0.30	0.46	0.94	0.312	0.50	0.39	0.55	0.91
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	0.029	0.31	0.31	0.40	0.91	0.411	0.65	0.50	0.70	0.85
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	0.035	0.33	0.33	0.38	0.89	0.277	0.58	0.51	0.77	0.93
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	0.014	0.30	0.30	0.49	0.95	0.595	0.77	0.49	0.86	0.87
$\hat{\mu}_{n,\text{SL}^{(1)}}$	0.032	0.32	0.32	0.38	0.90	0.131	0.41	0.39	0.40	0.87
$\hat{\mu}_{n,\text{SL}^{(2)}}$	0.037	0.33	0.33	0.37	0.88	0.102	0.41	0.40	0.40	0.87
$\hat{\mu}_{n,\text{SL}^{(3)}}$	0.001	0.30	0.30	0.46	0.94	0.155	0.39	0.36	0.53	0.92

NOTE: EST, estimator; BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; ASSE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; OR, outcome regression; PS, propensity score. No convergence for $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ was attained in five of the 1000 runs for the settings OR correct, PS correct ($\gamma = 2$) and OR incorrect, PS correct ($\gamma = 2$) and in three of the 1000 runs for the settings OR correct, PS incorrect and OR incorrect, PS incorrect.

Table 5.4: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 1,* $n = 1000$.

EST	BIAS	RMSE	MCSD	ASSE	COV	BIAS	RMSE	MCSD	ASSE	COV
					$n = 1000$					
	OR correct, PS correct ($\gamma = 1$)					OR incorrect, PS correct ($\gamma = 1$)				
$\hat{\mu}_{n,\text{MLE}}$	0.003	0.06	0.06	0.06	0.96	0.003	0.15	0.15	0.16	0.95
$\hat{\mu}_{n,\text{BR}}$	0.003	0.06	0.06	0.06	0.96	−0.003	0.09	0.09	0.09	0.94
$\hat{\mu}_{n,\text{SL}}$	0.005	0.06	0.06	–	–	−0.024	0.07	0.07	–	–
$\hat{\mu}_{n,\text{TMLE}}$	0.003	0.06	0.06	0.06	0.96	−0.004	0.12	0.12	0.13	0.95
$\hat{\mu}_{n,\text{SL-TMLE}}$	0.003	0.06	0.06	0.06	0.96	−0.003	0.07	0.07	0.07	0.93
$\hat{\mu}_{n,\text{MLE},(1)}$	0.004	0.06	0.06	0.06	0.95	0.013	0.09	0.09	0.08	0.93
$\hat{\mu}_{n,\text{MLE},(2)}$	0.005	0.06	0.06	0.06	0.95	0.014	0.10	0.10	0.09	0.93
$\hat{\mu}_{n,\text{MLE},(3)}$	0.003	0.06	0.06	0.06	0.96	−0.011	0.10	0.10	0.09	0.92
$\hat{\mu}_{n,\text{SL},(1)}$	0.004	0.06	0.06	0.06	0.95	0.000	0.07	0.07	0.07	0.94
$\hat{\mu}_{n,\text{SL},(2)}$	0.005	0.06	0.06	0.06	0.94	0.001	0.07	0.07	0.07	0.94
$\hat{\mu}_{n,\text{SL},(3)}$	0.003	0.06	0.06	0.06	0.95	−0.003	0.07	0.07	0.07	0.93
	OR correct, PS correct ($\gamma = 2$)					OR incorrect, PS correct ($\gamma = 2$)				
$\hat{\mu}_{n,\text{MLE}}$	−0.000	0.09	0.09	0.08	0.94	−0.041	0.50	0.49	0.35	0.78
$\hat{\mu}_{n,\text{BR}}$	0.000	0.08	0.08	0.08	0.91	−0.052	0.14	0.13	0.10	0.80
$\hat{\mu}_{n,\text{SL}}$	0.006	0.07	0.07	–	–	−0.113	0.15	0.10	–	–
$\hat{\mu}_{n,\text{TMLE}}$	0.004	0.09	0.09	0.08	0.90	0.044	0.19	0.19	0.13	0.85
$\hat{\mu}_{n,\text{SL-TMLE}}$	0.004	0.09	0.09	0.07	0.90	−0.012	0.10	0.10	0.08	0.87
$\hat{\mu}_{n,\text{MLE},(1)}$	0.027	0.12	0.12	0.07	0.78	0.168	0.30	0.24	0.13	0.66
$\hat{\mu}_{n,\text{MLE},(2)}$	0.030	0.13	0.12	0.07	0.78	0.152	0.27	0.23	0.11	0.64
$\hat{\mu}_{n,\text{MLE},(3)}$	0.001	0.08	0.08	0.08	0.91	−0.107	0.22	0.19	0.13	0.64
$\hat{\mu}_{n,\text{SL},(1)}$	0.025	0.12	0.12	0.07	0.78	0.038	0.13	0.12	0.08	0.79
$\hat{\mu}_{n,\text{SL},(2)}$	0.029	0.13	0.12	0.07	0.77	0.027	0.14	0.13	0.08	0.76
$\hat{\mu}_{n,\text{SL},(3)}$	0.001	0.08	0.08	0.07	0.91	−0.038	0.11	0.10	0.08	0.82
	OR correct, PS incorrect					OR incorrect, PS incorrect				
$\hat{\mu}_{n,\text{MLE}}$	−0.026	0.52	0.52	0.49	0.99	7.239	9.64	6.37	4.10	0.39
$\hat{\mu}_{n,\text{BR}}$	−0.006	0.11	0.11	0.11	0.94	1.237	1.28	0.33	0.28	0.01
$\hat{\mu}_{n,\text{SL}}$	−0.000	0.11	0.12	–	–	0.059	0.16	0.15	–	–
$\hat{\mu}_{n,\text{TMLE}}$	−0.004	0.12	0.12	0.41	1.00	1.222	1.26	0.29	0.66	0.37
$\hat{\mu}_{n,\text{SL-TMLE}}$	−0.005	0.12	0.12	0.39	1.00	0.100	0.18	0.15	0.42	1.00
$\hat{\mu}_{n,\text{MLE},(1)}$	−0.001	0.12	0.12	0.37	1.00	0.601	0.66	0.26	0.83	0.97
$\hat{\mu}_{n,\text{MLE},(2)}$	−0.001	0.12	0.12	0.36	1.00	0.573	0.63	0.25	0.80	1.00
$\hat{\mu}_{n,\text{MLE},(3)}$	0.003	0.16	0.16	0.42	1.00	0.289	0.46	0.36	0.85	0.97
$\hat{\mu}_{n,\text{SL},(1)}$	0.000	0.13	0.13	0.36	1.00	0.033	0.14	0.14	0.36	1.00
$\hat{\mu}_{n,\text{SL},(2)}$	0.001	0.12	0.12	0.36	1.00	0.035	0.14	0.14	0.35	1.00
$\hat{\mu}_{n,\text{SL},(3)}$	0.003	0.15	0.15	0.40	1.00	0.003	0.16	0.16	0.44	1.00

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; ASSE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; OR, outcome regression; PS, propensity score.

precision for the estimators that are based on the super-learner. Furthermore, it appears desirable to use a weighted loss-function (as $\hat{\mu}^{(3)}_{n,\text{MLE}}$ and $\hat{\mu}^{(3)}_{n,\text{SL}}$ do) as compared to fluctuation models using covariates with the estimated missingness

probabilities in the denominator (as for $\hat{\mu}^{(1)}_{n,\text{MLE}}$, $\hat{\mu}^{(2)}_{n,\text{MLE}}$, $\hat{\mu}^{(1)}_{n,\text{SL}}$ and $\hat{\mu}^{(2)}_{n,\text{SL}}$). In the case where the propensity score model is wrong but the outcome model is correct, all estimators show very similar behavior, except for the standard doubly robust estimator $\hat{\mu}_{n,\text{MLE}}$ which performs poorly. Finally, when both working models are misspecified, then $\hat{\mu}_{n,\text{BR}}$ drastically outperforms $\hat{\mu}_{n,\text{MLE}}$ in terms of bias and precision, as promised by the theory of bias-reduced doubly robust estimation. The proposed estimator $\hat{\mu}^{(3)}_{n,\text{SL}}$ performs best.

In conclusion, $\hat{\mu}^{(3)}_{n,\text{SL}}$ tends to perform best in Scenario 1, both in terms of bias and precision, especially when the inverse probability weights become highly variable. Adding the additional covariate $H^{(2)}$ to the fluctuation model for $\hat{\mu}^{(2)}_{n,\text{MLE}}$ and $\hat{\mu}^{(2)}_{n,\text{SL}}$ guarantees a substitution estimator, but does not lead to enhanced performance as compared to $\hat{\mu}^{(1)}_{n,\text{MLE}}$ and $\hat{\mu}^{(1)}_{n,\text{SL}}$. This can be understood upon noting that $\hat{\mu}^{(1)}_{n,\text{MLE}}$ and $\hat{\mu}^{(1)}_{n,\text{SL}}$ already have the form of an imputation estimator (see equation (5.20)).

Table 5.3 and Table 5.4 also show results on the performance of the sandwich standard error calculated using formula (5.21). The proposed estimator of the sandwich standard error performs well, especially for $\hat{\mu}^{(3)}_{n,\text{SL}}$, under correct specification of both working models, as well as when the outcome model is misspecified but the propensity score model is correct, and where the weights are not highly variable ($\gamma = 1$). When the weights become highly variable ($\gamma = 2$), the performance is worse. This is not a surprise because convergence to the normal limit distribution then happens more slowly. When the propensity model is misspecified (for both a correctly specified and misspecified outcome model), the sandwich standard errors overestimate the finite-sample variability of the estimator for all different estimators (except for $\hat{\mu}_{n,\text{BR}}$).

5.5.3 Scenario 2: Kang and Schafer setting

Data-generating mechanism

This simulation study is taken from Kang and Schafer (2007a) (see also Section 4.4.2) and often used as a benchmark to evaluate doubly robust estimators for the population mean outcome explainable by measured auxiliary covariates. For each

Table 5.5: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 2,* $n = 200$.

EST	BIAS	RMSE	MCSD	ASSE	COV	BIAS	RMSE	MCSD	ASSE	COV
	observed outcome RY					observed outcome $(1-R)Y$				
					$n = 200$					
	OR correct, PS correct					OR correct, PS correct				
$\hat{\mu}_{n,\text{MLE}}$	0.099	2.53	2.53	2.57	0.96	0.085	2.53	2.53	2.57	0.96
$\hat{\mu}_{n,\text{BR}}$	0.090	2.54	2.54	2.57	0.96	0.095	2.54	2.55	2.57	0.95
$\hat{\mu}_{n,\text{SL}}$	0.013	2.52	2.52	–	–	0.254	2.55	2.54	–	–
$\hat{\mu}_{n,\text{TMLE}}$	0.028	2.53	2.53	2.56	0.96	0.243	2.55	2.54	2.54	0.95
$\hat{\mu}_{n,\text{SL-TMLE}}$	0.028	2.53	2.53	2.56	0.96	0.243	2.55	2.54	2.54	0.95
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	−0.112	2.73	2.73	2.56	0.94	0.326	2.72	2.70	2.55	0.94
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	−0.153	2.80	2.80	2.57	0.94	0.348	2.77	2.75	2.56	0.94
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	0.029	2.53	2.53	2.56	0.96	0.241	2.55	2.54	2.54	0.95
$\hat{\mu}_{n,\text{SL}^{(1)}}$	−0.114	2.73	2.72	2.56	0.94	0.324	2.71	2.69	2.55	0.94
$\hat{\mu}_{n,\text{SL}^{(2)}}$	−0.186	2.80	2.80	2.57	0.94	0.365	2.77	2.75	2.56	0.94
$\hat{\mu}_{n,\text{SL}^{(3)}}$	0.025	2.53	2.53	2.56	0.96	0.244	2.55	2.54	2.54	0.95
	OR incorrect, PS incorrect					OR incorrect, PS incorrect				
$\hat{\mu}_{n,\text{MLE}}$	−15.150	88.60	87.34	15.92	0.93	4.759	6.05	3.74	3.31	0.60
$\hat{\mu}_{n,\text{BR}}$	−2.239	4.45	3.85	2.95	0.82	3.440	4.63	3.10	2.73	0.73
$\hat{\mu}_{n,\text{SL}}$	−1.912	3.68	3.15	–	–	4.742	5.83	3.40	–	–
$\hat{\mu}_{n,\text{TMLE}}$	−6.840	8.44	4.94	4.01	0.57	4.195	5.40	3.40	2.95	0.65
$\hat{\mu}_{n,\text{SL-TMLE}}$	−3.221	4.81	3.57	2.89	0.75	3.706	5.05	3.43	2.68	0.66
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	−5.962	7.83	5.09	3.05	0.50	1.447	4.02	3.75	2.90	0.84
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	−6.478	8.41	5.37	3.07	0.46	1.614	4.23	3.91	2.91	0.83
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	−3.263	5.23	4.09	3.28	0.80	3.656	4.77	3.07	2.66	0.70
$\hat{\mu}_{n,\text{SL}^{(1)}}$	−3.722	6.08	4.81	2.88	0.65	2.157	4.16	3.56	2.76	0.80
$\hat{\mu}_{n,\text{SL}^{(2)}}$	−3.944	6.41	5.05	2.89	0.64	2.293	4.38	3.73	2.77	0.79
$\hat{\mu}_{n,\text{SL}^{(3)}}$	−1.970	3.61	3.02	2.67	0.84	3.350	4.61	3.17	2.57	0.71

NOTE: EST, estimator; BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; ASSE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; OR, outcome regression; PS, propensity score. No convergence for $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ was attained in 13 of the 1000 runs for the settings OR correct, PS correct, $n = 200$ for both the observed outcome RY and $(1-R)Y$ and in five of the 1000 runs for the setting OR incorrect, PS incorrect, $n = 200$ for the observed outcome RY.

individual i $(i = 1, \ldots, n)$,

$$\boldsymbol{Z}_i \overset{d}{=} \mathbf{N}(\mathbf{0}, \mathbf{I}),$$
$$R_i | \boldsymbol{Z}_i \overset{d}{=} \text{Ber}\{g_0(\boldsymbol{Z}_i)\} \text{ and}$$
$$Y_i | \boldsymbol{Z}_i \overset{d}{=} N\{\bar{Q}_0(\boldsymbol{Z}_i), 1\},$$

where $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$, $\mathbf{I}$ is the 4×4 identity matrix, $g_0(\boldsymbol{Z}) = \text{expit}(-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4)$ and $\bar{Q}_0(\boldsymbol{Z}) = 210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4$. Misspecified working models are linear for the outcome model and logistic for the propensity score model, with covariates $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ with

$$
\begin{aligned}
X_1 &= \exp(Z_1/2), \\
X_2 &= Z_2/\{1 + \exp(Z_1)\} + 10, \\
X_3 &= (Z_1 Z_3/25 + 0.6)^3 \text{ and} \\
X_4 &= (Z_2 + Z_4 + 20)^2.
\end{aligned}
$$

The target parameter $E(Y) = \mu_0$ equals 210. In all cases, the marginal probability $P(R = 0)$ of the outcome Y being missing equals 0.5. We limit ourselves to the realistic settings where the working models both use either the covariates Z_k or the covariates X_k ($k = 1, \ldots, 4$) and thus both working models are correctly specified or both working models are incorrectly specified. We will show simulation results for two scenarios where either $R = 1$ or $R = 0$ denotes the data that are observed and results are shown in Table 5.5 ($n = 200$) and Table 5.6 ($n = 1000$).

Results

When both working models are correctly specified, all estimators have comparable performance, especially for $n = 1000$. However, $\hat{\mu}^{(1)}_{n,\text{MLE}}$, $\hat{\mu}^{(2)}_{n,\text{MLE}}$, $\hat{\mu}^{(1)}_{n,\text{SL}}$ and $\hat{\mu}^{(2)}_{n,\text{SL}}$ tend to be slightly more variable at $n = 200$, again illustrating the enhanced performance of using a weighted loss-function for estimators $\hat{\mu}^{(3)}_{n,\text{MLE}}$ and $\hat{\mu}^{(3)}_{n,\text{SL}}$. When both working models are misspecified, as in Kang and Schafer (2007a), $\hat{\mu}_{n,\text{MLE}}$ shows severe erratic behavior when the observed outcome RY is used; this behavior is partially eliminated when $(1-R)Y$ is used as the observed outcome (Robins et al. 2007). Confirming the theory of the bias-reduced doubly robust estimator from Section 5.3, $\hat{\mu}_{n,\text{BR}}$ does not show this severe erratic behavior (for both observed outcomes RY and $(1-R)Y$). Interestingly, the new estimator $\hat{\mu}^{(3)}_{n,\text{SL}}$ outperforms all existing estimators $\hat{\mu}_{n,\text{MLE}}$, $\hat{\mu}_{n,\text{BR}}$, $\hat{\mu}_{n,\text{TMLE}}$ and $\hat{\mu}_{n,\text{SL-TMLE}}$ for both observed outcomes RY and $(1-R)Y$. Do note that when $(1-R)Y$ is used as the observed outcome, $\hat{\mu}^{(1)}_{n,\text{MLE}}$ and $\hat{\mu}^{(1)}_{n,\text{SL}}$ show the best performance under double

Table 5.6: *Simulation results based on* 1000 *Monte Carlo replications for Scenario 2,* $n = 1000$.

EST	BIAS	RMSE	MCSD	ASSE	COV	BIAS	RMSE	MCSD	ASSE	COV
	observed outcome RY					observed outcome $(1-R)Y$				
					$n = 1000$					
	OR correct, PS correct					OR correct, PS correct				
$\hat{\mu}_{n,\text{MLE}}$	0.023	1.12	1.12	1.15	0.96	0.022	1.13	1.13	1.15	0.96
$\hat{\mu}_{n,\text{BR}}$	0.022	1.12	1.12	1.15	0.96	0.024	1.13	1.13	1.15	0.96
$\hat{\mu}_{n,\text{SL}}$	−0.006	1.13	1.13	–	–	0.070	1.13	1.12	–	–
$\hat{\mu}_{n,\text{TMLE}}$	0.013	1.12	1.12	1.15	0.96	0.062	1.13	1.13	1.14	0.96
$\hat{\mu}_{n,\text{SL-TMLE}}$	0.011	1.12	1.12	1.15	0.96	0.063	1.13	1.13	1.14	0.96
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	0.008	1.12	1.12	1.15	0.96	0.070	1.14	1.13	1.14	0.95
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	0.005	1.12	1.12	1.15	0.96	0.072	1.14	1.14	1.14	0.95
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	0.013	1.12	1.12	1.15	0.96	0.063	1.13	1.13	1.14	0.95
$\hat{\mu}_{n,\text{SL}^{(1)}}$	0.001	1.12	1.12	1.15	0.96	0.072	1.13	1.14	1.14	0.95
$\hat{\mu}_{n,\text{SL}^{(2)}}$	−0.001	1.12	1.12	1.15	0.96	0.078	1.14	1.14	1.14	0.95
$\hat{\mu}_{n,\text{SL}^{(3)}}$	0.011	1.12	1.12	1.15	0.96	0.064	1.13	1.13	1.14	0.95
	OR incorrect, PS incorrect					OR incorrect, PS incorrect				
$\hat{\mu}_{n,\text{MLE}}$	−53.715	469.04	466.19	49.69	0.78	4.509	4.83	1.73	1.60	0.19
$\hat{\mu}_{n,\text{BR}}$	−3.208	3.63	1.69	1.34	0.38	2.970	3.25	1.34	1.25	0.35
$\hat{\mu}_{n,\text{SL}}$	−2.879	3.18	1.52	–	–	3.356	3.74	1.65	–	–
$\hat{\mu}_{n,\text{TMLE}}$	−6.242	6.81	2.73	3.72	0.39	4.075	4.35	1.52	1.40	0.19
$\hat{\mu}_{n,\text{SL-TMLE}}$	−2.281	2.75	1.53	1.53	0.62	2.833	3.20	1.48	1.15	0.35
$\hat{\mu}_{n,\text{MLE}^{(1)}}$	−3.878	4.55	2.39	1.98	0.51	1.442	2.35	1.86	1.35	0.73
$\hat{\mu}_{n,\text{MLE}^{(2)}}$	−3.738	4.47	2.45	1.92	0.52	1.570	2.46	1.90	1.34	0.70
$\hat{\mu}_{n,\text{MLE}^{(3)}}$	−3.351	4.86	1.92	2.82	0.74	3.110	3.39	1.34	1.23	0.30
$\hat{\mu}_{n,\text{SL}^{(1)}}$	−2.771	3.38	1.94	1.28	0.46	1.919	2.38	1.42	1.17	0.60
$\hat{\mu}_{n,\text{SL}^{(2)}}$	−3.001	3.60	1.99	1.28	0.41	1.981	2.46	1.46	1.17	0.58
$\hat{\mu}_{n,\text{SL}^{(3)}}$	−2.188	2.52	1.24	1.21	0.55	2.571	2.91	1.37	1.13	0.41

NOTE: EST, estimator; BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; ASSE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; OR, outcome regression; PS, propensity score.

model misspecification.

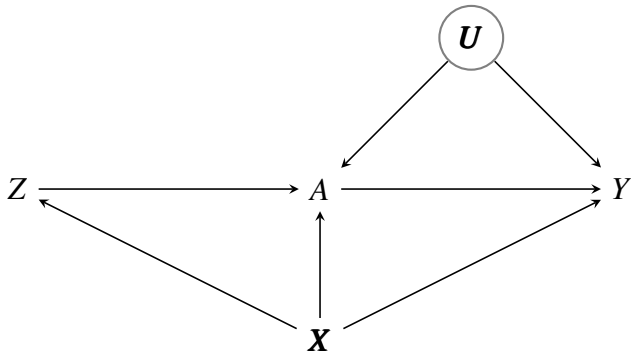
Table 5.5 and Table 5.6 also show the performance of the sandwich standard errors (5.21). Excellent finite-sample behavior is seen when both working models are correctly specified. When both working models are misspecified, the approximation of the sandwich standard errors is better for $n = 1000$ than for $n = 200$, but not sufficient, as is the case for TMLE. Unlike in Scenario 1, we do observe that they underestimate the true finite-sample variability of the estimators. For the proposed estimators, the best performance is seen for $\hat{\mu}^{(3)}_{n,\text{SL}}$.

5.6 Linear Instrumental Variable Analyses

In this section, we show that the proposed procedure is not restricted to the doubly robust estimator from Section 5.2 but can be easily extended to other doubly robust estimators. We will do this by illustrating how the principle can be implemented for linear instrumental variable analyses.

5.6.1 Doubly robust estimation of linear instrumental variable models

Suppose we are interested in the causal effect of an exposure A on an outcome Y in the presence of unmeasured confounding $\boldsymbol{U}$ and in the presence of an instrumental variable (IV) Z (Robins 1994; Wooldridge 2002; Hernán and Robins 2006). In particular, we assume that for measured covariates $\boldsymbol{X}$, (a) Z is associated with A conditional on $\boldsymbol{X}$, (b) $Z \perp\!\!\!\perp Y|A,\boldsymbol{X},\boldsymbol{U}$ and (c) $Z \perp\!\!\!\perp \boldsymbol{U}|\boldsymbol{X}$. The unmeasured variables $\boldsymbol{U}$ are thus such that $(\boldsymbol{X}^T,\boldsymbol{U}^T)^T$ would be sufficient to control for confounding of the causal effect of A on Y. These assumptions can be visualized by means of the following causal diagram:



The observed data is given by the i.i.d. sample $\mathcal{O} = (\boldsymbol{O}_1,\ldots,\boldsymbol{O}_n)$ of size n with P_0 the unknown underlying data-generating mechanism of the observables $\boldsymbol{O} = (Y,A,Z,\boldsymbol{X})$, where we consider both A and Z to be dichotomous. We consider inference for the causal effect τ_0 indexing the linear instrumental variable model

$$E(Y|A,Z,\boldsymbol{X},\boldsymbol{U}) = E(Y|A=0,Z,\boldsymbol{X},\boldsymbol{U}) + \tau_0 A. \tag{5.22}$$

Let $\mathcal{M}$ denote the statistical model for P_0 implied by assumptions (a)-(b)-(c) and (5.22). Okui et al. (2012) show that a doubly robust estimator of τ_0 can be obtained by solving

$$0 = \sum_{i=1}^{n} \{Z_i - \hat{g}_n(\boldsymbol{X}_i)\} \left\{ Y_i - \tau A_i - \hat{\bar{Q}}_n(\boldsymbol{X}_i) \right\}. \tag{5.23}$$

Here, $\hat{g}_n(\boldsymbol{X})$ is an estimator of the conditional mean of the instrument given the measured confounders $\boldsymbol{X}$, $g_0(\boldsymbol{X}) = E(Z|\boldsymbol{X})$, based on a nuisance working model $\mathcal{G} = \{g(\boldsymbol{X})|g \text{ in some class of functions}\}$. Let $g^*(\boldsymbol{X})$ denote the probability limit ($\hat{g}_n(\boldsymbol{X}) \xrightarrow{p} g^*(\boldsymbol{X})$) such that $g^*(\boldsymbol{X}) = g_0(\boldsymbol{X})$ when $\mathcal{M}(\mathcal{G})$ holds, where $\mathcal{M}(\mathcal{G})$ represents the statistical model for the joint distribution of $\boldsymbol{O}$ implied by the working model $\mathcal{G}$. Next, $\hat{\bar{Q}}_n(\boldsymbol{X})$ is an estimator of the conditional mean of the outcome given the measured confounders $\boldsymbol{X}$ among the non-exposed, $\bar{Q}_0(\boldsymbol{X}) = E(Y|A=0,\boldsymbol{X})$, based on a nuisance working model $\mathcal{Q} = \{\bar{Q}(\boldsymbol{X})|\bar{Q} \text{ in some class of functions}\}$. Let $\bar{Q}^*(\boldsymbol{X})$ denote the probability limit ($\hat{\bar{Q}}_n(\boldsymbol{X}) \xrightarrow{p} \bar{Q}^*(\boldsymbol{X})$) such that $\bar{Q}^*(\boldsymbol{X}) = \bar{Q}_0(\boldsymbol{X})$ when $\mathcal{M}(\mathcal{Q})$ holds, where $\mathcal{M}(\mathcal{Q})$ represents the statistical model for the joint distribution of $\boldsymbol{O}$ implied by the working model $\mathcal{Q}$. Consistency of the doubly robust estimator $\hat{\tau}_{n,\mathrm{DR}}(\hat{g}_n, \hat{\bar{Q}}_n)$ is then attained under the model $\mathcal{M} \cap \{\mathcal{M}(\mathcal{Q}) \cup \mathcal{M}(\mathcal{G})\}$. At the intersection model $\mathcal{M} \cap \mathcal{M}(\mathcal{Q}) \cap \mathcal{M}(\mathcal{G})$, it has the following expansion

$$\hat{\tau}_{n,\mathrm{DR}}(\hat{g}_n, \hat{\bar{Q}}_n) - \tau_0 = (P_n - P_0)\{D^*_{\mathrm{IV}}(g_0, \bar{Q}_0; \tau_0)\} + o_p(n^{-1/2}), \tag{5.24}$$

with influence function $D^*_{\mathrm{IV}}(g_0, \bar{Q}_0; \tau_0)$ given by

$$D^*_{\mathrm{IV}}(g_0, \bar{Q}_0; \tau_0)(\boldsymbol{O}) = \frac{\{Z - g_0(\boldsymbol{X})\}\{Y - \bar{Q}_0(\boldsymbol{X})\}}{E\left[A\{Z - g_0(\boldsymbol{X})\}\right]} - \tau_0. \tag{5.25}$$

Note that by construction,

$$P_n\left[D^*_{\mathrm{IV}}\left\{\hat{g}_n, \hat{\bar{Q}}_n; \hat{\tau}_{n,\mathrm{DR}}(\hat{g}_n, \hat{\bar{Q}}_n)\right\}\right] = 0.$$

5.6.2 Practical implementation of the proposed procedure

The proposed (data-adaptive) bias-reduced doubly robust procedure works as follows:

Step 1: Estimator $\hat{g}_n^{\text{MLE}}$ for the conditional mean $g_0(\boldsymbol{X})$ of the instrument. Postulate a parametric working model for $g_0(\boldsymbol{X})$: $\mathcal{G}_{\boldsymbol{\psi}} = \{g(\boldsymbol{\psi})(\boldsymbol{X}) | \boldsymbol{\psi} \in \mathbb{R}^s\}$, where $g(\boldsymbol{\psi})(\boldsymbol{X}) = G\{\boldsymbol{\psi}^T \boldsymbol{l}(\boldsymbol{X})\}$ and G an appropriate inverse link function, e.g., $G(\cdot) = \text{expit}(\cdot)$, and $\boldsymbol{l} = (1, l_1, \ldots, l_{s-1})$. Obtain the MLE $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ using the estimating function $D_g^{\text{MLE}}(\boldsymbol{\psi})(\boldsymbol{O})$ given in (5.5) but with R replaced by Z. Let $\hat{g}_n^{\text{MLE}} = g(\hat{\boldsymbol{\psi}}_n^{\text{MLE}})$.

Step 2: Initial estimator $\hat{\bar{Q}}_n^0$ for the conditional mean outcome among the non-exposed $\bar{Q}_0$. We again consider two options: (1) Postulate a parametric working model for $\bar{Q}_0(\boldsymbol{X})$: $\mathcal{Q}_{\boldsymbol{\xi}} = \{\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) | \boldsymbol{\xi} \in \mathbb{R}^r\}$, where $\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) = Q\{\boldsymbol{\xi}^T \boldsymbol{k}(\boldsymbol{X})\}$ and Q an appropriate inverse link function and $\boldsymbol{k} = (1, k_1, \ldots, k_{r-1})$. Obtain the MLE $\hat{\boldsymbol{\xi}}_n^{\text{MLE}}$ using the estimating function $D_Q^{\text{MLE}}(\boldsymbol{\xi})(\boldsymbol{O})$ given in (5.6) but with R replaced by $1-A$. Let $\hat{\bar{Q}}_n^{\text{MLE}} = \bar{Q}(\hat{\boldsymbol{\xi}}_n^{\text{MLE}})$; (2) Use a super-learner based on a library $\{\hat{\bar{Q}}_j | j = 1, \ldots, J\}$ to obtain a fit $\hat{\bar{Q}}_n^{\text{SL}}$ among those with $A = 0$. Let the initial estimator $\hat{\bar{Q}}_n^0$ denote either $\hat{\bar{Q}}_n^{\text{MLE}}$ or $\hat{\bar{Q}}_n^{\text{SL}}$.

Step 3: Fluctuation $\hat{\bar{Q}}_n^1$ of the initial estimator $\hat{\bar{Q}}_n^0$. Following the same argument as in Step 3 of Section 5.4.2, we assume that $Y \in [0,1]$. In this step, we fluctuate the initial estimator $\hat{\bar{Q}}_n^0$ by means of a parametric fluctuation model $\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon})$ such that the score with respect to $\boldsymbol{\varepsilon}$ guarantees bias reduction in the direction $\boldsymbol{\psi}$. For this purpose, define the fluctuation model $\left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon}) : \boldsymbol{\varepsilon} \in \mathbb{R}^s\right\}$ through the initial estimator ($\hat{\bar{Q}}_n^0(\boldsymbol{0}) = \hat{\bar{Q}}_n^0$):

$$\text{logit}\, \hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon})(\boldsymbol{X}) = \text{logit}\, \hat{\bar{Q}}_n^0(\boldsymbol{X}) + \boldsymbol{\varepsilon}^T \widehat{\boldsymbol{H}}(\hat{g}_n^{\text{MLE}})(Z, \boldsymbol{X}), \tag{5.26}$$

with

$$\begin{aligned} \widehat{\boldsymbol{H}}(\hat{g}_n^{\text{MLE}})(Z, \boldsymbol{X}) &= G'\{\hat{\boldsymbol{\psi}}_n^{\text{MLE},T} \boldsymbol{l}(\boldsymbol{X})\} \boldsymbol{l}(\boldsymbol{X}) + [Z - G\{\hat{\boldsymbol{\psi}}_n^{\text{MLE},T} \boldsymbol{l}(\boldsymbol{X})\}] \widehat{\boldsymbol{W}}(\hat{g}_n^{\text{MLE}}), \\ \widehat{\boldsymbol{W}}(\hat{g}_n^{\text{MLE}}) &= \frac{n^{-1} \sum_{j=1}^n A_j G'\{\hat{\boldsymbol{\psi}}_n^{\text{MLE},T} \boldsymbol{l}(\boldsymbol{X}_j)\} \boldsymbol{l}(\boldsymbol{X}_j)}{n^{-1} \sum_{j=1}^n A_j [Z_j - G\{\hat{\boldsymbol{\psi}}_n^{\text{MLE},T} \boldsymbol{l}(\boldsymbol{X}_j)\}]}. \end{aligned}$$

Then define $\hat{\boldsymbol{\varepsilon}}_n = \arg\min_{\boldsymbol{\varepsilon}} P_n \left[\mathcal{L}^{(1)} \left\{\hat{\bar{Q}}_n^0(\boldsymbol{\varepsilon})\right\}\right]$, with $\mathcal{L}^{(1)}$ the quasi-log-likelihood loss-function (5.13). This can be obtained via standard logistic regression of the outcome on the covariates $\widehat{\boldsymbol{H}}$ using as offset $\text{logit}\, \hat{\bar{Q}}_n^0$. Next, define the updated

estimator as $\hat{\bar{Q}}_n^1 = \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n)$. By construction of the fluctuation model, it follows that $\hat{\boldsymbol{\varepsilon}}_n$ solves

$$P_n\left\{D^*_{\boldsymbol{\psi},\text{IV}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^1)\right\} = \mathbf{0},$$

with $D^*_{\boldsymbol{\psi},\text{IV}}\left\{g(\boldsymbol{\psi}), \bar{Q}\right\} = \partial D^*_{\text{IV}}\left\{g(\boldsymbol{\psi}), \bar{Q}; \tau_0\right\}/\partial\boldsymbol{\psi}$, implying bias reduction in the direction of $\boldsymbol{\psi}$.

Step 4: Estimating the target parameter $\hat{\tau}_n$. Given the estimators $\hat{g}_n^{\text{MLE}}$ and $\hat{\bar{Q}}_n^1$, we obtain the doubly robust estimator $\hat{\tau}_n \equiv \hat{\tau}_{n,\text{DR}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^1)$.

5.7 Discussion

In this chapter, we proposed an extension to the bias-reduced doubly robust estimation principle, originally proposed in Vermeulen and Vansteelandt (2015a) and discussed in Chapter 4, which we refer to as data-adaptive bias-reduced doubly robust estimation. In particular, we relaxed the restriction to parametric nuisance working models by making use of data-adaptive learning algorithms to estimate the conditional mean outcome. We carry on to work with a parametric working model for the missingness mechanism for the inferential problem of Section 5.2 and with a parametric working model for the conditional distribution of the IV in Section 5.6 in order to avoid positivity violations. Because the concern of bias is greater for such parametric models, we applied the bias-reduction principle of Section 5.3.2 (see also Chapter 4) in the direction of the nuisance parameters indexing these parametric working models. As a side-effect, the two nuisance working models need not be of the same dimension, unlike for the original bias-reduced estimation principle. We furthermore illustrated that the proposed extension is not restricted to doubly robust estimation of the mean outcome susceptible to missingness explainable by measured covariates but also extends to other doubly robust procedures as illustrated in Section 5.6.

This new procedure follows the spirit of the TMLE procedure of van der Laan and Rubin (2006) in the sense it also extends an initial data-adaptive estimator of the relevant part $\bar{Q}_0(\boldsymbol{X})$ of the conditional outcome distribution, enhancing the performance of the estimator of the target parameter. Fluctuation of initial

parametric estimators is also seen in other contexts, e.g., Tan (2010); Rotnitzky et al. (2012).

In the implementation of Section 5.4.2, we could additionally have considered a fluctuation $\hat{g}_n^{\text{MLE}}(\boldsymbol{\varepsilon}_g)$ of the initial estimator $\hat{g}_n^{\text{MLE}}$ for the missingness mechanism in a way to obtain bias-reduction in the direction of the fluctuation parameters $\boldsymbol{\varepsilon}$ used in the construction of the fluctuation model $\hat{\tilde{Q}}_n^0(\boldsymbol{\varepsilon})$. This, however, would demand iterating the proposed procedure to ensure that the probability limits $g^*(\boldsymbol{X})$ and $\hat{\tilde{Q}}^*(\boldsymbol{X})$ of the final estimators locally minimize the squared first-order asymptotic bias of the doubly robust estimator. In simulation studies however, we observed this algorithm to be unstable and non-convergent.

A limitation of our proposal is that the current asymptotic linearity theorem, presented in Appendix 5.A, only guarantees valid inference under a correct working model for the propensity score. This is because misspecification of the propensity score working model (but correct specification of the outcome model) would demand acknowledging the uncertainty of the estimator for the outcome working model so as to make the remainder term R_n in the expansion (5.2) of second-order to obtain valid inference. It is not clear how to accomplish this when using data-adaptive learning algorithms. In Appendix 5.A, we show how this can be done when a parametric working model for the conditional mean outcome is used. An alternative for standard error calculation, which does not demand the assumption of a correctly specified missingness model, is to use the nonparametric bootstrap, which however lacks supporting theory when the estimators rely on data-adaptive estimation (van der Laan 2014). van der Laan (2014) suggests one option to obtain valid inference under misspecification of one of both working models in the context of the TMLE procedure. This is accomplished by additionally fluctuating initial estimators of the working models such that the scores with respect to the corresponding fluctuation parameters guarantee the remainder term in the expansion (5.2) to be of second-order so as to accomplish asymptotic linearity. In the context of the current paper, this would demand adding an additional covariate to the fluctuation models considered in Section 5.4.2. For details, we refer to van der Laan (2014), see also Section 4.3.3.

5.A Asymptotic Linearity Theorem

We present an asymptotic linearity theorem with corresponding influence function for the doubly robust estimators $\hat{\mu}_n^{(c)}$ ($c = 1,2,3$) of Section 5.4 under the assumption of a correctly specified working model for the propensity score $\mathcal{M}(\mathcal{G})$ but a potentially misspecified working model for the conditional mean outcome $\mathcal{M}(\mathcal{Q})$. The derivation below relies on empirical process theory, for which we refer to van der Vaart and Wellner (1996); Gill (1989). A summary of the key concepts is given in the Appendix A.1 of van der Laan and Rose (2011).

We let g^*_{MLE} denote the probability limit of $\hat{g}_n^{\text{MLE}}$ (the propensity score estimator) and let $\bar{Q}^*_{(c)}$ denote the probability limit of $\hat{\bar{Q}}_n^{(c)}$ (the updated estimator for the conditional mean outcome); thus $\hat{g}_n^{\text{MLE}} \xrightarrow{p} g^*_{\text{MLE}}$ and $\hat{\bar{Q}}_n^{(c)} \xrightarrow{p} \bar{Q}^*_{(c)}$. Because we assume a correctly specified model for the missingness mechanism, we have that $g^*_{\text{MLE}} = g_0$. This implies that

$$P_0\{D^*(g^*_{\text{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\} = 0.$$

By definition of the proposed estimation strategy, we have that (for $c = 1,2,3$)

$$\begin{aligned} P_n\left\{D^*(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(c)}; \hat{\mu}_n^{(c)})\right\} &= 0, \\ P_n\left\{D_g^{\text{MLE}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}})\right\} &= \mathbf{0}, \\ P_n\left\{D^*_{\boldsymbol{\psi}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(c)})\right\} &= \mathbf{0}. \end{aligned}$$

We make the following regularity conditions;

Donsker class condition: Suppose that the set $\mathcal{D}_1 = \left\{D^*(g, \bar{Q}; \mu_0) : (g, \bar{Q})\right\}$ is a P_0-Donsker class where $(g, \bar{Q})$ varies over a set containing the sequences $(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(c)})$ and $(\hat{g}_n^{\text{MLE}}, \bar{Q}^*_{(c)})$ with probability tending to one.

Consistency condition of D^*: Assume that

$$P_0\left[\left\{D^*(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(g^*_{\text{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\}^2\right] \rightarrow 0,$$

$$P_0\left[\left\{D^*(\hat{g}_n^{\mathrm{MLE}},\bar{Q}_{(c)}^*;\mu_0)-D^*(g_{\mathrm{MLE}}^*,\bar{Q}_{(c)}^*;\mu_0)\right\}^2\right] \;\to\; 0,$$

in probability as $n\to\infty$.

Glivenko-Cantelli class condition: Suppose that $\mathcal{D}_2=\left\{D_{\boldsymbol{\psi}}^*(g,\bar{Q}):(g,\bar{Q})\right\}$ is a P_0-Glivenko-Cantelli class where $(g,\bar{Q})$ varies over a set containing the sequence $(\hat{g}_n^{\mathrm{MLE}},\hat{\bar{Q}}_n^{(c)})$ with probability tending to one.

Consistency condition of $D_{\boldsymbol{\psi}}^*$: With $D_{\boldsymbol{\psi}}^*=(D_{\boldsymbol{\psi},1}^*,\ldots,D_{\boldsymbol{\psi},s}^*)^T$, assume that

$$P_0\left[\left\{D_{\boldsymbol{\psi},i}^*(\hat{g}_n^{\mathrm{MLE}},\hat{\bar{Q}}_n^{(c)})-D_{\boldsymbol{\psi},i}^*(g_{\mathrm{MLE}}^*,\bar{Q}_{(c)}^*)\right\}^2\right]\to 0$$

in probability as $n\to\infty$ for $i=1,\ldots,s$.

Second-order term condition: Define the second order term

$$R_n^*=P_0\left[R\left\{\hat{\bar{Q}}_n^{(c)}(\boldsymbol{X})-\bar{Q}_{(c)}^*(\boldsymbol{X})\right\}\left\{\frac{\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X})-g_{\mathrm{MLE}}^*(\boldsymbol{X})}{\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X})g_{\mathrm{MLE}}^*(\boldsymbol{X})}\right\}\right]$$

and assume $R_n^*=o_p(n^{1/2})$. Note that this second-order term involves the product of the differences $\hat{\bar{Q}}_n^{(c)}(\boldsymbol{X})-\bar{Q}_{(c)}^*(\boldsymbol{X})$ and $\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X})-g_{\mathrm{MLE}}^*(\boldsymbol{X})$.

Theorem 5.1 (Asymptotic linearity of the proposed doubly robust estimator)**.** *Assuming the regularity conditions given above, we have that the doubly robust estimator $\hat{\mu}_n^{(c)}$ ($c=1,2,3$), constructed in Section 5.4, is an asymptotically linear estimator of μ_0 at P_0 with influence function $D^*(g_0,\bar{Q}_{(c)}^*;\mu_0)$. That is,*

$$\hat{\mu}_n^{(c)}-\mu_0=(P_n-P_0)D^*(g_0,\bar{Q}_{(c)}^*;\mu_0)+o_p(n^{-1/2}).$$

In particular, $n^{1/2}(\hat{\mu}_n^{(c)}-\mu_0)$ converges in distribution to a normal distribution with mean zero and variance $\sigma_0^2=P_0[\{D^*(g_0,\bar{Q}_{(c)}^*;\mu_0)\}^2]$.

Proof. By definition of the proposed estimator, we know that

$$P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \hat{\mu}_n^{(c)})\right\} = 0.$$

Furthermore, because we assume the model for the missingness mechanism to be correctly specified, we have that $P_0\{D^*(g^*_{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\} = 0$. From this, it follows that

$$\begin{aligned}
\hat{\mu}_n^{(c)} - \mu_0 &= n^{-1}\sum_{i=1}^{n}\left[\frac{R_i}{\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X}_i)}\left\{Y_i - \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i)\right\} + \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}_i) - \mu_0\right] \\
&= P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0)\right\} \\
&= P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0)\right\} - P_0\left\{D^*(g^*_{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \\
&= (P_n - P_0)\left\{D^*(g^*_{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \\
&\quad + P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(g^*_{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \\
&= (P_n - P_0)\left\{D^*(g^*_{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \\
&\quad + P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \qquad (5.27) \\
&\quad + P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0) - D^*(g^*_{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\}. \qquad (5.28)
\end{aligned}$$

Let us first consider the term (5.27). We have

$$\begin{aligned}
&P_n\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \\
&\quad = (P_n - P_0)\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} \\
&\quad\quad + P_0\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\}
\end{aligned}$$

and from the Donsker class condition and the consistency condition of D^*, it follows that

$$(P_n - P_0)\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\} = o_p(n^{-1/2}).$$

Next consider the term $P_0\left\{D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\right\}$. We have

that

$$
\begin{aligned}
P_0 \Big\{ D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}_{(c)}^*; \mu_0) \Big\} \\
&= P_0 \Big\{ D^*(g_{\mathrm{MLE}}^*, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(g_{\mathrm{MLE}}^*, \bar{Q}_{(c)}^*; \mu_0) \Big\} \\
&\quad + P_0 \Big\{ D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}_{(c)}^*; \mu_0) \\
&\qquad - D^*(g_{\mathrm{MLE}}^*, \hat{\bar{Q}}_n^{(c)}; \mu_0) + D^*(g_{\mathrm{MLE}}^*, \bar{Q}_{(c)}^*; \mu_0) \Big\}.
\end{aligned}
$$

Because

$$
\begin{aligned}
P_0 \Big\{ & D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}_{(c)}^*; \mu_0) \\
& - D^*(g_{\mathrm{MLE}}^*, \hat{\bar{Q}}_n^{(c)}; \mu_0) + D^*(g_{\mathrm{MLE}}^*, \bar{Q}_{(c)}^*; \mu_0) \Big\} \\
&= P_0 \left(\frac{R}{g_{\mathrm{MLE}}^*(\boldsymbol{X})} \left[\{Y - \bar{Q}_{(c)}^*(\boldsymbol{X})\} - \{Y - \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X})\} \right] \right. \\
&\qquad \left. - \frac{R}{\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X})} \left[\{Y - \bar{Q}_{(c)}^*(\boldsymbol{X})\} - \{Y - \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X})\} \right] \right) \\
&= P_0 \left[R \left\{ \frac{1}{g_{\mathrm{MLE}}^*(\boldsymbol{X})} - \frac{1}{\hat{g}_n^{\mathrm{MLE}}(\boldsymbol{X})} \right\} \left\{ \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}) - \bar{Q}_{(c)}^*(\boldsymbol{X}) \right\} \right] \\
&= R_n^*,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
P_0 \Big\{ & D^*(\hat{g}_n^{\mathrm{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(\hat{g}_n^{\mathrm{MLE}}, \bar{Q}_{(c)}^*; \mu_0) \Big\} \\
&= P_0 \Big\{ D^*(g_{\mathrm{MLE}}^*, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(g_{\mathrm{MLE}}^*, \bar{Q}_{(c)}^*; \mu_0) \Big\} + R_n^*
\end{aligned}
$$

and by assumption, the second-order term $R_n^* = o_p(n^{-1/2})$. Finally, note that

$$
\begin{aligned}
P_0 \Big\{ & D^*(g_{\mathrm{MLE}}^*, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(g_{\mathrm{MLE}}^*, \bar{Q}_{(c)}^*; \mu_0) \Big\} \\
&= -P_0 \left[\left\{ \hat{\bar{Q}}_n^{(c)}(\boldsymbol{X}) - \bar{Q}_{(c)}^*(\boldsymbol{X}) \right\} \left\{ \frac{R - g_{\mathrm{MLE}}^*(\boldsymbol{X})}{g_{\mathrm{MLE}}^*(\boldsymbol{X})} \right\} \right],
\end{aligned}
$$

which is $o_p(n^{-1/2})$ under the assumption that $g_{\mathrm{MLE}}^*(\boldsymbol{X})$ equals the true missingness

mechanism $g_0(\boldsymbol{X})$. We conclude that (5.27) is $o_p(n^{-1/2})$.

We next consider the second term (5.28), which is the contribution of the estimation of the propensity score. From the way we estimated the fluctuation parameter $\boldsymbol{\varepsilon}^{(c)}$, it will follow that we can ignore this term; that is, (5.28) is $o_p(n^{-1/2})$, even under misspecification of the estimator of $\bar{Q}_0$. Define the function $\Upsilon_{(c)}(\boldsymbol{\psi}) = D^*\{g(\boldsymbol{\psi}), \bar{Q}^*_{(c)}; \mu_0\}$. It follows that (5.28) equals $P_n\{\Upsilon_{(c)}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}) - \Upsilon_{(c)}(\boldsymbol{\psi}_0)\}$. We know that $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ is an asymptotically linear estimator for $\boldsymbol{\psi}_0$ with influence function $-P_0\left\{D^{\text{MLE}}_{\boldsymbol{\psi},g}(\boldsymbol{\psi}_0)\right\} D^{\text{MLE}}_g(\boldsymbol{\psi}_0)$, that is,

$$\hat{\boldsymbol{\psi}}_n^{\text{MLE}} - \boldsymbol{\psi}_0 = (P_n - P_0)\left[-P_0\left\{D^{\text{MLE}}_{\boldsymbol{\psi},g}(\boldsymbol{\psi}_0)\right\}^{-1} D^{\text{MLE}}_g(\boldsymbol{\psi}_0)\right] + o_p(n^{-1/2})$$

and where $D^{\text{MLE}}_{\boldsymbol{\psi},g}(\boldsymbol{\psi}_0) = \partial D^{\text{MLE}}_g(\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$. Consider the Taylor expansion

$$\Upsilon_{(c)}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}) - \Upsilon_{(c)}(\boldsymbol{\psi}_0) = \Upsilon^T_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)(\hat{\boldsymbol{\psi}}_n^{\text{MLE}} - \boldsymbol{\psi}_0) + o_p(|\hat{\boldsymbol{\psi}}_n^{\text{MLE}} - \boldsymbol{\psi}_0|)$$

with $\Upsilon_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0) = \partial\Upsilon_{(c)}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$. Note that by the asymptotic linearity of $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$, $o_p(|\hat{\boldsymbol{\psi}}_n^{\text{MLE}} - \boldsymbol{\psi}_0|) = o_p(n^{-1/2})$. Consequently,

$$\begin{aligned}
&P_n\left\{\Upsilon_{(c)}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}) - \Upsilon_{(c)}(\boldsymbol{\psi}_0)\right\} \\
&\quad = P_n\left\{\Upsilon^T_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)\right\}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}} - \boldsymbol{\psi}_0) + o_p(n^{-1/2}) \\
&\quad = (P_n - P_0)\left[-P_0\left\{\Upsilon^T_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)\right\} P_0\left\{D^{\text{MLE}}_{\boldsymbol{\psi},g}(\boldsymbol{\psi}_0)\right\}^{-1} D^{\text{MLE}}_g(\boldsymbol{\psi}_0)\right] + o_p(n^{-1/2}),
\end{aligned}$$

where the last equality follows from the asymptotic linearity of $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ and the fact that $P_n\left\{\Upsilon^T_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)\right\} = P_0\left\{\Upsilon^T_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)\right\} + o_p(n^{-1/2})$. In principle, when the probability limit $\bar{Q}^*_{(c)}$ is different from the truth $\bar{Q}_0$, this term would contribute to the influence function of $\hat{\mu}_n^{(c)}$. However, by definition of $\hat{\boldsymbol{\varepsilon}}_n^{(c)}$ and since that $\hat{\bar{Q}}_n^{(c)} = \hat{\bar{Q}}_n^0(\hat{\boldsymbol{\varepsilon}}_n^{(c)})$, we have that $P_n\left\{D^*_{\boldsymbol{\psi}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(c)})\right\} = \mathbf{0}$ for every n. From the Glivenko-Cantelli class condition and the consistency condition of $D^*_{\boldsymbol{\psi}}$, it then follows that $P_n\{D^*_{\boldsymbol{\psi}}(\hat{g}_n^{\text{MLE}}, \hat{\bar{Q}}_n^{(c)})\} \xrightarrow{p} P_0\{D^*_{\boldsymbol{\psi}}(g^*_{\text{MLE}}, \bar{Q}^*_{(c)})\} = P_0\{\Upsilon_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)\}$. Hence, it follows that $P_0\{\Upsilon_{\boldsymbol{\psi},(c)}(\boldsymbol{\psi}_0)\} = \mathbf{0}$. We conclude that (5.28) is $o_p(n^{-1/2})$.

Putting everything together, we find that

$$\hat{\mu}_n^{(c)} - \mu_0 = (P_n - P_0)\left\{D^*(g_0, \bar{Q}^*_{(c)}; \mu_0)\right\} + o_p(n^{-1/2}),$$

showing asymptotic linearity of $\hat{\mu}_n^{(c)}$. □

Remark 5.3. *When the updated estimator $\hat{\bar{Q}}_n^{(c)}$ ($c = 1,2,3$) is based on a parametric model for the initial estimator, the Donsker class condition will be satisfied. Furthermore, when the updated estimator $\hat{\bar{Q}}_n^{(c)}$ ($c = 1,2,3$) is based on the super-learner $\hat{\bar{Q}}_n^{SL}$ as an initial estimator and each of the estimators in the library falls in a Donsker class, the Donsker class condition will also be satisfied for $\hat{\bar{Q}}_n^{SL}$ because the convex combination of such estimators also falls in that class (van der Vaart and Wellner 1996). In short, the Donsker class condition is satisfied if it holds for each of the estimators in the library of the super-learner. Examples for such estimators are for instance given in van der Laan (2014).*

Theorem 5.1 establishes asymptotic linearity of the doubly robust estimators $\hat{\mu}_n^{(c)}$ ($c = 1,2,3$) under model $\mathcal{M}(\mathcal{G})$. To obtain asymptotic linearity under model $\mathcal{M}(\mathcal{Q})$ (so that $\bar{Q}^*_{(c)} = \bar{Q}_0$), one should assume asymptotic linearity of $\hat{\bar{Q}}_n^{(c)}$ in the sense that $P_0\{D^*(g^*_{\text{MLE}}, \hat{\bar{Q}}_n^{(c)}; \mu_0) - D^*(g^*_{\text{MLE}}, \bar{Q}^*_{(c)}; \mu_0)\} = (P_n - P_0)D^*_{\bar{Q}^*_{(c)}} + o_p(n^{-1/2})$ (see van der Laan and Rose (2011), p. 572 for a discussion concerning this assumption). In this case, the influence function of $\hat{\mu}_n^{(c)}$ under model $\mathcal{M}(\mathcal{Q})$ becomes $D^*(g^*_{\text{MLE}}, \bar{Q}^*_{(c)}; \mu_0) + D^*_{\bar{Q}^*_{(c)}}$. When the initial estimator $\hat{\bar{Q}}_n^0$ for the conditional mean outcome $\bar{Q}_0$ is parametric such as $\bar{Q}(\boldsymbol{\xi})(\boldsymbol{X}) = Q\{\boldsymbol{\xi}^T \boldsymbol{l}(\boldsymbol{X})\}$, and $\boldsymbol{\xi}$ is estimated via a root-n consistent estimator such as the MLE, asymptotic linearity of $\hat{\mu}_n^{(c)}$ ($c = 1,2,3$) under model $\mathcal{M}(\mathcal{Q})$ can be shown in a straightforward manner, as we argue next. In this case, the updated estimator $\hat{\bar{Q}}_n^{(c)}$ is based on a parametric model $\bar{Q}(\boldsymbol{\theta}^{(c)})$ for the conditional mean outcome with dimension equal to the dimension of $\boldsymbol{\theta}^{(c)} = (\boldsymbol{\xi}^T, \boldsymbol{\varepsilon}^{(c),T})^T$. Let $\hat{\boldsymbol{\theta}}_n^{(c)}$ denote the corresponding estimator of $\boldsymbol{\theta}^{(c)}$ with probability limit $\boldsymbol{\theta}^*_{(c)}$. When the parametric model for the conditional mean outcome is correctly specified, $\boldsymbol{\theta}^*_{(c)} = (\boldsymbol{\xi}_0^T, \mathbf{0}^T)^T$. Under standard regularity

conditions, $\hat{\boldsymbol{\theta}}_n^{(c)}$ is asymptotically linear with influence function $D^*_{\bar{Q}}(\boldsymbol{\theta}^{(c)})$; that is, $(\hat{\boldsymbol{\theta}}_n^{(c)} - \boldsymbol{\theta}^*_{(c)}) = (P_n - P_0)D^*_{\bar{Q}}(\boldsymbol{\theta}^*_{(c)}) + o_p(n^{-1/2})$. We now show that under $\mathcal{M}(\mathcal{Q})$, (5.27) is not $o_p(n^{-1/2})$ because then $g^*_{\text{MLE}} \neq g_0$. For this purpose, define the function $\Theta_{(c)}(\boldsymbol{\theta}^{(c)}, \boldsymbol{\psi}) = D^*\{g(\boldsymbol{\psi}), \bar{Q}(\boldsymbol{\theta}^{(c)}); \mu_0\}$. It follows from a Taylor expansion and the asymptotic linearity of $\hat{\boldsymbol{\theta}}_n^{(c)}$, that (5.27) equals

$$\begin{aligned}
&P_n\left\{\Theta_{(c)}(\hat{\boldsymbol{\theta}}_n^{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}}) - \Theta_{(c)}(\boldsymbol{\theta}^*_{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}})\right\} \\
&\quad = P_n\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}})\right\}(\hat{\boldsymbol{\theta}}_n^{(c)} - \boldsymbol{\theta}^*_{(c)}) + o_p(|\hat{\boldsymbol{\theta}}_n^{(c)} - \boldsymbol{\theta}^*_{(c)}|) \\
&\quad = P_n\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}})\right\}(P_n - P_0)D^*_{\bar{Q}}(\boldsymbol{\theta}^*_{(c)}) + o_p(n^{-1/2}),
\end{aligned}$$

with $\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}}) = \partial\Theta_{(c)}(\boldsymbol{\theta}^{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}})/\partial\boldsymbol{\theta}^{(c)}|_{\boldsymbol{\theta}^{(c)}=\boldsymbol{\theta}^*_{(c)}}$. Under suitable regularity conditions (Robins et al. (1994), app. B), it follows from the uniform weak law of large numbers that $P_n\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \hat{\boldsymbol{\psi}}_n^{\text{MLE}})\right\}$ converges in probability to $P_0\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \boldsymbol{\psi}^*_{\text{MLE}})\right\}$. This shows that (5.27) can be written as

$$(P_n - P_0)\left[P_0\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \boldsymbol{\psi}^*_{\text{MLE}})\right\}D^*_{\bar{Q}}(\boldsymbol{\theta}^*_{(c)})\right] + o_p(n^{-1/2}).$$

Upon replacing $\boldsymbol{\psi}_0$ by $\boldsymbol{\psi}^*_{\text{MLE}}$ in the remainder of the proof of Theorem 5.1, it follows that the influence function of $\hat{\mu}_n^{(c)}$ under model $\mathcal{M}(\mathcal{Q})$ equals

$$D^*(g^*_{\text{MLE}}, \bar{Q}^*_{(c)}; \mu_0) + P_0\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \boldsymbol{\psi}^*_{\text{MLE}})\right\}D^*_{\bar{Q}}(\boldsymbol{\theta}^*_{(c)}).$$

For this specific setting, we thus have that

$$D^*_{\bar{Q}^*_{(c)}} = P_0\left\{\Theta^T_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \boldsymbol{\psi}^*_{\text{MLE}})\right\}D^*_{\bar{Q}}(\boldsymbol{\theta}^*_{(c)}).$$

Local-ancillarity with respect to $\boldsymbol{\xi}$ can however still be obtained by implementing bias-reduction in the direction of $\boldsymbol{\xi}$ as well, enforcing this term to be zero (see Section 5.3).

Remark 5.4. *Under model $\mathcal{M}(\mathcal{G})$, $\boldsymbol{\psi}^*_{MLE} = \boldsymbol{\psi}_0$ and hence $P_0\{\Theta_{\boldsymbol{\theta}^{(c)},(c)}(\boldsymbol{\theta}^*_{(c)}, \boldsymbol{\psi}_0)\} = \mathbf{0}$. We may thus conclude from the derivation*

above (but now with the probability limit $\boldsymbol{\theta}^*_{(c)}$ not necessarily equal to $(\boldsymbol{\xi}_0^T, \mathbf{0}^T)^T$) that (5.27) is $o_p(n^{-1/2})$ for a parametric initial estimator for $\bar{Q}_0$ which is potentially misspecified, given the parametric model for the propensity score is correctly specified. This argument is more insightful than the Donsker class condition and the second-order term condition.

5.B R-Functions

R-function

Below, we provide an R-function to obtain the data-adaptive bias-reduced doubly robust estimators $\hat{\mu}_n^{(c)}$ to estimate the mean outcome $E(Y)$ in the presence of incomplete data as outlined in Section 5.4 for both a linear regression working model and a super-learner working model for the initial estimator of the conditional mean outcome and with a logistic regression working model for the propensity score.

As input, the function uses the missingness indicator `R`, the outcome `Y`, the auxiliary covariates `cov`, the estimation method for the initial estimator of the conditional mean outcome `type.initQ=c("par","SL")` (either `"par"` for a parametric linear regression model or `"SL"` for a super-learner), the level `zeta` at which the initial estimator must be truncated, the type of loss-function for the logistic fluctuation model `fluc=c("unweighted","weighted")`, and the level of significance `alpha` of a hypothesis test of the mean equal to `mu.tilde`. As output, the function delivers the estimate `est`, which equals $\hat{\mu}_n^{(c)}$, the standard error `se`, which equals $\widehat{\text{SE}}(\hat{\mu}_n^{(c)})$ based on equation (5.21), a $(1-\alpha)100\%$ Wald confidence interval `ci`, the Wald statistic `Wald.statistic` and the corresponding p-value `p.value` for a test of the null hypothesis $H_0 : \mu = \tilde{\mu}$. We also provide an example where the procedure is applied to a random dataset, obtained via the Kang and Schafer data-generating mechanism of Section 5.5.3.

```
data.adaptive.biasreduced.DR<-
          function(R,Y,cov,type.initQ=c("par","SL"),
          zeta=0.005,fluc=c("unweighted","weighted"),
```

```
        alpha=0.05,mu.tilde=0){
expit <- function(x){1/(1+exp(-x))}
logit <- function(x){log(x/(1-x))}
n <- length(R)
dat.cov <- data.frame(cov)
int.cov <- cbind(rep(1,n),cov)
colnames(dat.cov) <-
          paste("cov.",1:dim(cov)[2],sep="")

# propensity score
mler <- glm(R~cov,family="binomial")
ps.par <- predict(mler,type="response")

# initial conditional mean outcome
a <- min(Y[R==1]) - 0.1*abs(min(Y[R==1]))
b <- max(Y[R==1]) + 0.1*abs(max(Y[R==1]))
Y.star <- (Y-a)/(b-a)
{
if(type.initQ=="par"){
  mley <- lm(Y.star~cov,subset=(R==1))
  initQ <- predict(mley,newdata=dat.cov)
}
else if(type.initQ=="SL"){
  Ym.star <- Y.star[R==1]
  dat.cov.m <- dat.cov[R==1,,drop=FALSE]
  SL.library <- c("SL.glm", "SL.randomForest",
        "SL.gam","SL.polymars", "SL.mean")
  initQ <- SuperLearner(Y=Ym.star,X=dat.cov.m,
        newX=dat.cov,verbose = FALSE,
        SL.library=SL.library,
        method="method.NNLS")$SL.predict
}
}
```

```
initQ.trunc <- ifelse(initQ<zeta,zeta,
        ifelse(initQ>1-zeta,1-zeta,initQ))

# fluctuation
{
if(fluc=="unweighted"){
  w.cov <- (1-ps.par)/ps.par*int.cov
  fluctuationQ <- glm(Y.star~-1+w.cov,
        family=binomial,offset=logit(initQ.trunc),
        subset=(R==1))
  flucQ <- expit(logit(initQ.trunc)+
  as.vector(coef(fluctuationQ)%*%t(w.cov)))
}
else if(fluc=="weighted"){
  fluctuationQ <-glm(Y.star~cov,
        family=binomial,offset=logit(initQ.trunc),
        subset=(R==1),weights=(1-ps.par)/ps.par)
  flucQ <- expit(logit(initQ.trunc)+
        as.vector(coef(fluctuationQ)%*%t(int.cov)))
}
}

# doubly robust estimator
U<-function(R,Y,outcome,ps){
  outcome+R/ps*(Y-outcome)
}
est.trunc <- mean(U(R=R,Y=Y.star,
        outcome=flucQ,ps=ps.par))
mu <- (b-a)*est.trunc+a

# standard error
se.mu <- (b-a)*sd(U(R=R,Y=Y.star,
        outcome=flucQ,ps=ps.par))/sqrt(n)
```

```
  # 95% confidence interval
  ci.mu <- mu+c(-1,1)*qnorm(1-alpha/2)*se.mu

  # Wald test statistic
  W <- (mu-mu.tilde)/se.mu

  # p-value Wald test
  p.value <- 2*pnorm(abs(W),lower.tail=FALSE)

  return(list(est=mu,se=se.mu,ci=ci.mu,
          Wald.statistic=W,p.value=p.value))
}
```


Example

```
library(SuperLearner)

gen.dataKS <- function(k,n,mech=
          c("normal","reverse"),spec=c("C","I")){
  set.seed(k)
  z1 <- rnorm(n);z2 <- rnorm(n)
  z3 <- rnorm(n);z4 <- rnorm(n)
  z <- cbind(z1,z2,z3,z4)
  colnames(z) <- paste("Z.",1:4,sep="")
  x1 <- exp(0.5*z1)
  x2 <- z2/(1+exp(z1))+10
  x3 <- (0.04*z1*z3+0.6)^3
  x4 <- (z2+z4+20)^2
  y <- rnorm(n,210+27.4*z1+13.7*z2+13.7*z3+
          13.7*z4,sd=1)
  r <- rbinom(n,1,expit(-1.5*z1+0.75*z2-
          0.375*z3-0.15*z4))
```

```
  if(spec=="C"){
    x1 <- z1;x2 <- z2
    x3 <- z3;x4 <- z4;
    x <- z
  }
  if(mech=="reverse"){
    r <- 1-r
  }
  data <- data.frame(r,y,x1,x2,x3,x4)
  names(data) <- c("r","y","x.1","x.2","x.3","x.4")
  data
}

data <- gen.dataKS(1,n=1000,mech="reverse",spec="I")
R <- data$r
Y <- data$y
cov <- cbind(data$x.1,data$x.2,data$x.3,data$x.4)

dataadaptive.biasreduced.DR(R=R,Y=Y,cov=cov,
          type.initQ="SL",zeta=0.005,
          fluc="weighted",alpha=0.05,mu.tilde=210)
```

CHAPTER 6

Increasing the Power of the Mann-Whitney Test in Randomized Experiments Through Flexible Covariate Adjustment

The Mann-Whitney U test is frequently used to evaluate treatment effects in randomized experiments with skewed outcome distributions or small sample sizes. It may lack power, however, because it ignores the auxiliary baseline covariate information that is routinely collected. Wald and score tests in so-called Probabilistic Index Models (PIMs) generalize the Mann-Whitney U test to enable adjustment for covariates, but these may lack robustness by demanding correct model specification and do not lend themselves to small sample inference. Using semiparametric efficiency theory, we here propose an alternative extension of the Mann-Whitney U test which increases its power by exploiting covariate information in an objective way and which lends itself to permutation inference. Simulation studies and an application to an HIV clinical trial show that the proposed permutation test attains the nominal Type I error rate and can be drastically more powerful than the classical Mann-Whitney U test.

6.1 Introduction

Randomized experiments are routinely performed to detect the effect of a novel treatment as compared to placebo or standard treatment. For continuous outcomes, it is common practice to evaluate the treatment effect via a standard two-sample t-test, which evaluates a difference in mean outcomes between treated and untreated subjects. When outcome distributions are heavy tailed or the sample size is small, the nonparametric Mann-Whitney U test (Mann and Whitney 1947) (or equivalently the Wilcoxon rank-sum test (Wilcoxon 1945)) is often used, instead. The corresponding effect size measure is the so-called **marginal probabilistic index** (MPI): the probability that a randomly chosen treated subject has a higher outcome than a randomly chosen untreated subject.

In many randomized experiments, extensive baseline data are collected for each subject prior to treatment assignment, such as baseline outcome data, data on medical history, demographic data, etc. Both the two-sample t-test and the Mann-Whitney U test ignore this covariate information and may therefore lack power as compared to analyses that involve covariate adjustment. Traditionally, covariate adjustment is performed via regression, resulting in conditional effect sizes. In particular, covariate adjustment of the Mann-Whitney U test can be accomplished via probabilistic index models (PIMs) (Thas et al. 2012; Brumback et al. 2006), which model the probability that the outcome of a randomly chosen treated subject is higher than the outcome of a randomly chosen untreated subject, in function of the covariates. While such adjustment may increase the power to detect a treatment effect, extensive debate exists as to whether this is appropriate (Hauck et al. 1998; Lewis 1999; Assmann et al. 2000; Raab et al. 2000; Senn 2000; Pocock et al. 2002; Grouin et al. 2004). A first reason for this concern is the post hoc selection of covariates: the fact that covariate adjustment may prompt **fishing expeditions** of those covariates that yield the largest estimate or the smallest p-value for the treatment effect (Pocock et al. 2002). It has therefore been argued that covariate adjustment should only be considered under pre-specified models with respect to covariates that are pre-specified in the study protocol (Hauck et al. 1998; Grouin et al. 2004). This is difficult at the design stage when the associations between outcome and covariates are not yet well understood (Pocock et al. 2002). A second

reason for this debate is that one may obtain biased estimates of the treatment effect when the association between covariates and outcome is misspecified and thus the model assumptions fail to hold (Rosenblum and van der Laan 2009). Third, whether covariate adjustment increases power is a subtle question in nonlinear models. On the one hand, unlike in linear models, adding baseline covariates to a nonlinear regression model that are independent of the treatment, may increase the variability of the treatment effect estimate (Robinson and Jewell 1991). On the other hand, adding such baseline covariates to nonlinear models also changes the magnitude of the treatment effect (Greenland et al. 1999). In particular, by non-collapsibility (Greenland et al. 1999) of nonlinear effect measures, conditional effect sizes tend to deviate more from the null hypothesis of no effect than the corresponding marginal effect size. A related point of discussion is whether to target a marginal effect size or a conditional effect size. A marginal effect size may be of main interest in the primary analysis of randomized experiments since it does not demand modeling assumptions and because the primary interest then lies in an overall effect to enable policy making. Conditional effect sizes may be of special interest in secondary analyses if one is interested in detecting whether the treatment is particularly useful or harmful in certain subpopulations or because they may be better transportable across populations (Vansteelandt and Keiding 2011). In this chapter, we mainly focus on marginal effect sizes.

To resolve the above concerns, Tsiatis et al. (2008) (see also Leon et al. (2003)) developed a strategy that allows for principled and flexible covariate adjustment based on a semiparametric efficient estimator of the marginal treatment mean difference. Their approach separates evaluation of the treatment difference from covariate adjustment, thus allowing for objective and optimal exploitation of covariate-outcome associations. This idea has been extended to more than two treatments (Zhang et al. 2008) and more general outcome summaries such as odds ratios (Zhang et al. 2008; Moore and van der Laan 2009). In this chapter, we also take a semiparametric theory perspective to implement covariate adjustment by identifying an efficient estimator for the MPI, the effect size considered by the Mann-Whitney U statistic.

In Section 6.2, we formalize the problem and define the marginal treatment effect measure of interest. In Section 6.3, we describe a regression framework to

adjust for baseline covariates on the scale of the considered treatment effect (Thas et al. 2012), discuss its potential shortcomings and show in Section 6.4 how this framework can be used to obtain an estimator of the MPI. Section 6.5 presents semiparametric theory results and considers the practical implementation of the proposed locally efficient estimator of the MPI. This estimator has asymptotic variance equal to the semiparametric variance bound if a working model for the unknown joint distribution of outcomes and baseline covariates is correct, but remains consistent if that model is misspecified. Finally, in Section 6.6, we develop a permutation test of the null hypothesis of no treatment effect and describe how this can be implemented. An application of the proposed methods to an HIV clinical trial is presented in Section 6.7 and simulation studies in Section 6.8 illustrate its desirable performance relative to other methods.

6.2 The MPI in Randomized Experiments

Consider a randomized experiment where n subjects are randomly drawn from some population and randomized to an experimental treatment ($A = 1$) or to a standard treatment or placebo ($A = 0$). Suppose that for every subject a $(p \times 1)$-dimensional vector of auxiliary baseline covariates $\boldsymbol{X}$ is available, measured prior to treatment assignment so that, by randomization, A is independent of $\boldsymbol{X}$, denoted $A \perp\!\!\!\perp \boldsymbol{X}$. Interest lies in the effect of the treatment A on an ordinal discrete or continuous outcome Y. The observed data are thus the i.i.d. (identically and independently distributed) sample $\mathcal{O} = \{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$ with $\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i)$.

The Mann-Whitney U test (Mann and Whitney 1947; Wilcoxon 1945) is frequently used for testing a treatment effect when the distribution of the outcome within treatment groups deviates from the normal distribution or when normality is difficult to assess due to the small sample size. The Mann-Whitney test statistic

$$U = \sum_{i=1}^{n} \sum_{j \neq i} (1 - A_i) A_j I(Y_i \preceq Y_j), \tag{6.1}$$

with $I(Y_i \preceq Y_j) = I(Y_i < Y_j) + 0.5 I(Y_i = Y_j)$ and $I(\cdot)$ the ordinary indicator function, is based on calculating the number of pairs for which the outcome of a treated

($A = 1$) subject is larger than the outcome of an untreated ($A = 0$) subject plus a half times the number of pairs for which the outcome of a treated subject equals the outcome of an untreated subject; the latter accounts for ties. Let $N_1 = \sum_{i=1}^{n} A_i$ and $N_0 = \sum_{i=1}^{n}(1 - A_i)$ denote the number of treated and untreated subjects which we allow to be random, in line with typical randomized designs. The proportion U/N_0N_1 then estimates the so called **Marginal Probabilistic Index** (MPI) (Acion et al. 2006)

$$\begin{aligned} \nu_0 &= P(Y \preceq Y^* | A = 0, A^* = 1) \\ &= P(Y < Y^* | A = 0, A^* = 1) + 0.5P(Y = Y^* | A = 0, A^* = 1), \end{aligned} \tag{6.2}$$

which encodes the probability that a randomly selected treated subject with data $(Y^*, A^* = 1, \boldsymbol{X}^*)$ has a higher outcome than a randomly selected untreated subject with data $(Y, A = 0, \boldsymbol{X})$, where both subjects are independent. This equals the so-called area under the receiver operator characteristic curve (AUC) comparing responses of two treatments (Grissom 1994; Hanley and McNeil 1982). For a continuous outcome, $P(Y = Y^* | A = 0, A^* = 1) = 0$, and hence the MPI equals $P(Y \preceq Y^* | A = 0, A^* = 1) = P(Y < Y^* | A = 0, A^* = 1)$. When there is no treatment effect, we expect the proportion U/N_0N_1 to be close to one half. When treatment is beneficial, where we assume higher outcome values are better, the proportion will tend to be higher than one half; when treatment is harmful, this proportion will tend to be lower than one half.

The Mann-Whitney U test thus comes with a useful effect size measure, the MPI, which is an attractive effect size measure because it maintains its meaning across a variety of outcome measures (such as continuous, ordinal or binary outcomes) and across a variety of distributions (such as skewed distributions in which case a mean difference may not constitute a meaningful effect size, see Acion et al. (2006)). The Mann-Whitney U test itself then aims to detect if the MPI (6.2) deviates from 0.5. Although the MPI is an attractive effect size measure who's use is advocated by many authors (such as in D'Agostino et al. (2006); Acion et al. (2006); Zhou (2008)), it can be easily misinterpreted, see for instance Hand (1992); Senn (2006, 2011, 2012). The MPI can indeed be easily misinterpreted as being the probability that a patient benefits from receiving the treatment rather then

not receiving the treatment while in fact, the MPI compares the outcomes of two independent randomly selected patients, of which one is treated and the other one is not treated.

When we only observe the i.i.d. data (Y_i, A_i) for $i = 1, \ldots, n$ and no auxiliary information is available, U/N_0N_1 is the best estimator for $P(Y \preceq Y^*|A = 0, A^* = 1)$ in the sense that it is consistent and asymptotically unbiased with minimum variance in finite samples under the nonparametric model for the distribution of Y given A (Lehmann 1951). In contrast, when auxiliary information is available, more efficient estimation of the MPI can be obtained by exploiting the known independence of A and $\boldsymbol{X}$. In this chapter, we will propose an efficient estimator for $P(Y \preceq Y^*|A = 0, A^* = 1)$ using the available auxiliary information. We will next use this efficient estimator to develop a distribution free test that is anticipated to be more powerful than the Mann-Whitney U test in small samples.

6

6.3 Model-Based Regression Adjustment

6.3.1 Standard regression adjustment

The marginal additive treatment effect defined by $\beta_0 = E(Y|A = 1) - E(Y|A = 0)$ for a continuous outcome forms the basis of the two-sample t-test. Like the MPI ν_0, the effect size β_0 is unconditional. To exploit covariate information, one may instead consider a conditional additive treatment effect $E(Y|A = 1, \boldsymbol{X} = \boldsymbol{x}) - E(Y|A = 0, \boldsymbol{X} = \boldsymbol{x})$ and estimate this under a linear regression model. For example, we may posit the model

$$E(Y|A, \boldsymbol{X}; \boldsymbol{\beta}) = \beta_{\text{int}} + \boldsymbol{\beta}_{\boldsymbol{X}}^T \boldsymbol{X} + \beta_A A \tag{6.3}$$

with $\boldsymbol{\beta} = (\beta_{\text{int}}, \boldsymbol{\beta}_{\boldsymbol{X}}^T, \beta_A)^T$, in which case the conditional additive treatment effect equals β_A, not depending on $\boldsymbol{x}$ and coinciding with β_0. Alternatively, we may posit a linear model allowing for a treatment-covariate interaction such as $E(Y|A, \boldsymbol{X}; \boldsymbol{\beta}) = \beta_{\text{int}} + \boldsymbol{\beta}_{\boldsymbol{X}}^T \boldsymbol{X} + \beta_A A + \boldsymbol{\beta}_{A\boldsymbol{X}}^T A\boldsymbol{X}$ with $\boldsymbol{\beta} = (\beta_{\text{int}}, \boldsymbol{\beta}_{\boldsymbol{X}}^T, \beta_A, \boldsymbol{\beta}_{A\boldsymbol{X}}^T)^T$, in which case the conditional additive treatment effect equals $\beta_A + \boldsymbol{\beta}_{A\boldsymbol{X}}^T \boldsymbol{x}$, which depends on the covariates. When the outcome is dichotomous, interest may lie in the

marginal log odds ratio $\beta_0 = \log\{\text{odds}_{A=1}(Y)/\text{odds}_{A=0}(Y)\}$, with $\text{odds}_{A=a}(Y) = P(Y=1|A=a)/P(Y=0|A=a)$. To exploit covariate information, one may then consider a conditional log odds ratio $\log\{\text{odds}_{A=1}(Y|\boldsymbol{X}=\boldsymbol{x})/\text{odds}_{A=0}(Y|\boldsymbol{X}=\boldsymbol{x})\}$ with $\text{odds}_{A=a}(Y|\boldsymbol{X}) = P(Y=1|A=a,\boldsymbol{X})/P(Y=0|A=a,\boldsymbol{X})$. This can be estimated under a logistic regression model. For example, we may posit the model

$$\text{logit}\,P(Y=1|A,\boldsymbol{X};\boldsymbol{\beta}) = \beta_{\text{int}} + \boldsymbol{\beta}_{\boldsymbol{X}}^T\boldsymbol{X} + \beta_A A, \tag{6.4}$$

with $\boldsymbol{\beta} = (\beta_{\text{int}}, \boldsymbol{\beta}_{\boldsymbol{X}}^T, \beta_A)^T$ and $\text{logit}(x) = \log\{x/(1-x)\}$, in which case the conditional log odds ratio equals β_A. Even when the model (6.4) is correctly specified (and thus no treatment-covariate interactions are present), a marginal and conditional log odds ratio β_0 and β_A are generally not identical due to non-collapsibility (Greenland et al. 1999).

6.3.2 Regression adjustment via PIMs

One may now likewise extend the MPI $P(Y \preceq Y^*|A=0,A^*=1)$ to a **Conditional Probabilistic Index** (CPI) $P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X}=\boldsymbol{x},\boldsymbol{X}^*=\boldsymbol{x})$. This encodes the probability that for two randomly chosen subjects with the same covariates, the treated subject has a higher outcome than the untreated subject; it thus incorporates covariate information. This can be estimated under a so called **Probabilistic Index Model** (PIM) (Thas et al. 2012; Brumback et al. 2006), such as

$$\text{logit}\,P(Y \preceq Y^*|A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}) = \tau_A(A^*-A) + \boldsymbol{\tau}_{\boldsymbol{X}}^T(\boldsymbol{X}^*-\boldsymbol{X}) \tag{6.5}$$

with $\boldsymbol{\tau} = (\tau_A, \boldsymbol{\tau}_{\boldsymbol{X}}^T)^T$. Under this model, one obtains a covariate-adjusted effect size $P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X}=\boldsymbol{x},\boldsymbol{X}^*=\boldsymbol{x};\boldsymbol{\tau}) = \text{expit}(\tau_A)$, with $\text{expit}(x) = e^x/(1+e^x)$. Treatment-covariate interactions can be allowed for by considering the model

$$\begin{aligned}&\text{logit}\,P(Y \preceq Y^*|A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})\\ &\quad= \tau_A(A^*-A) + \boldsymbol{\tau}_{\boldsymbol{X}}^T(\boldsymbol{X}^*-\boldsymbol{X}) + \tau_{A\boldsymbol{X}}^T(A^*\boldsymbol{X}^* - A\boldsymbol{X})\end{aligned} \tag{6.6}$$

with $\boldsymbol{\tau} = (\tau_A, \boldsymbol{\tau}_{\boldsymbol{X}}^T, \boldsymbol{\tau}_{A\boldsymbol{X}}^T)^T$, resulting in the conditional effect size $P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X}=\boldsymbol{x},\boldsymbol{X}^*=\boldsymbol{x};\boldsymbol{\tau}) = \text{expit}(\tau_A + \boldsymbol{\tau}_{A\boldsymbol{X}}^T\boldsymbol{x})$, which may not be constant across

covariate values. Models (6.5) and (6.6) can be fitted using the `pim` package for R available on R-forge (De Neve and Sabbe 2013). For a summary of parameter estimation and inference in PIMs, we refer to Appendix 6.A

Like the two-sample t-test fits into the linear regression framework, the classical Mann-Whitney U test can be expressed using a PIM. In Thas et al. (2012), this correspondence is shown using an identity link. Likewise, with a logit link, $\text{logit}\, P(Y \preceq Y^*|A,A^*;\tau) = \tau(A^* - A)$, the estimation strategy of Thas et al. (2012) yields the estimator $\hat{\nu}_n = \text{expit}(\hat{\tau}_n) = U/N_0N_1$ of ν_0, which involves the Mann-Whitney U statistic, see Appendix 6.B. Likewise, to test for the presence of a (covariate adjusted) treatment effect under model (6.5), one may simply use a Wald test of the null hypothesis that $\tau_A = 0$ based on a sandwich estimator. However, in small samples, such sandwich estimators may not well approximate the true sampling variability of the estimated treatment effect coefficient(s) (see the simulation results in Section 6.8). In our experience, the sandwich estimator is sometimes even impossible to calculate because of singularities due to matrix inversion in small samples. Given that the Mann-Whitney U test is often indicated in small sample settings, regression adjustment for auxiliary covariates using a PIM to improve power of randomized trial analyses is thus limiting. In the next sections, we propose an alternative strategy, which overcomes this concern and those listed in the introduction.

6.4 Standardization of the CPI

One way to construct an estimator for the MPI which includes covariate information, is to aggregate predictions from the PIM over the covariate distribution. This is known as **standardization** (Vansteelandt and Keiding 2011). It exploits the identity

$$E\{P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*)|A=0,A^*=1\} = P(Y \preceq Y^*|A=0,A^*=1)$$

to arrive at estimators of the MPI. In particular, given a PIM $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$ for the CPI $P(Y \preceq Y^*|A,A^*,\boldsymbol{X},\boldsymbol{X}^*)$ and an estimator $\hat{\boldsymbol{\tau}}_n$ of $\boldsymbol{\tau}$, an estimator for the MPI

can be calculated as

$$\hat{\nu}_{n,\text{IMP}} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n). \tag{6.7}$$

For instance, under model (6.5),

$$\hat{\nu}_{n,\text{IMP}} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \text{expit}\{\hat{\tau}_{n,A} + \hat{\boldsymbol{\tau}}_{n,\boldsymbol{X}}^T(\boldsymbol{X}_j - \boldsymbol{X}_i)\}.$$

The estimator $\hat{\nu}_{n,\text{IMP}}$ can be interpreted as a regression-imputation estimator for the MPI. Indeed, for every pair of individuals (i,j) with $i \neq j$, we predict the probability that the jth individual has a higher outcome than the ith individual, given their observed covariates $\boldsymbol{X}_j$ and $\boldsymbol{X}_i$, as if (potentially contrary to the fact) the jth individual were treated and the ith individual were not treated.

In general, $\hat{\nu}_{n,\text{IMP}}$ may be a biased estimator of ν_0 under misspecification of the PIM. This is worrisome, especially considering that unbiased estimation is attainable in randomized experiments. Interestingly, however, Theorem 6.1 below shows that certain fitting strategies for the PIM deliver an asymptotically unbiased estimator of the MPI, even when the PIM is misspecified. This is formalized in that $\hat{\nu}_{n,\text{IMP}}$ is asymptotically unbiased under model $\mathcal{M}_{\text{indep}}$, which is defined as the model that leaves the law of $\boldsymbol{O} = (Y,A,\boldsymbol{X})$ unrestricted except for the known independence of A and $\boldsymbol{X}$. Theorem 6.1 further specifies that correct modeling of the PIM delivers more efficient estimators of the MPI.

> **Theorem 6.1.** *When the CPI is modeled with a PIM that uses a logit link function, includes a main effect for the treatment (i.e., the PIM includes the term $\tau_A(A_j - A_i)$) and is fitted using the default estimation method in the* `pim` *R-package as detailed in equation (6.22), the estimator $\hat{\nu}_{n,IMP}$ is asymptotically unbiased for the MPI under model $\mathcal{M}_{indep}$, i.e., even under misspecification of the PIM. When the PIM is also correctly specified, (6.7) achieves the efficiency bound for CAN (consistent and asymptotically normal) estimators of the MPI under model $\mathcal{M}_{indep}$, because the estimator is* ***locally efficient****.*

Theorem 6.1 follows from the semiparametric theory results given in the sub-

sequent section, where we also give a proof. These results furthermore imply that asymptotic inference for $\hat{\nu}_{n,\text{IMP}}$ can be performed without the need for calculating estimators of the variance of the estimated regression coefficients $\hat{\boldsymbol{\tau}}_n$.

6.5 Semiparametric Inference: Augmentation of the Mann-Whitney U Test Statistic

In this section, we give the formal semiparametric theory that underlies Theorem 6.1 in Section 6.4. Specifically, we present an adaptation of the Mann-Whitney U test statistic that incorporates covariate information but which is not susceptible to bias due to model misspecification; it only assumes the validity of model $\mathcal{M}_{\text{indep}}$. The results follow the spirit of the work of Tsiatis et al. (2008) who introduced a strategy that allows for principled and flexible covariate adjustment based on a locally efficient estimator for the marginal additive treatment effect in model $\mathcal{M}_{\text{indep}}$. Their approach separates evaluation of the treatment difference from the covariate adjustment process, thereby allowing for objective exploitation of covariate-outcome associations.

6.5.1 Semiparametric inference

To develop the formal semiparametric efficiency theory for the estimation of the MPI ν_0 under the model $\mathcal{M}_{\text{indep}}$, we first derive the class of all CAN estimators for ν_0 under model $\mathcal{M}(\pi)$, which is similar to $\mathcal{M}_{\text{indep}}$, but additionally considers the randomization probability π known. Afterwards, we derive the class of all CAN estimators for ν_0 under model $\mathcal{M}_{\text{indep}}$. Finally, we identify the efficient CAN estimator for ν_0.

Class of all consistent and asymptotically normal estimators for ν_0 under model $\mathcal{M}(\pi)$

Model $\mathcal{M}(\pi)$ for the i.i.d. data $\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, defined by $P(A = 1|\boldsymbol{X}) = P(A = 1) = \pi$, $0 < \pi < 1$ known, can be formalized as the set of all joint

laws

$$\mathcal{M}(\pi) = \Big\{ f_{Y,A,\boldsymbol{X}}(y,a,\boldsymbol{x};\boldsymbol{\eta}) = f_{Y|A,\boldsymbol{X}}(y|a,\boldsymbol{x};\boldsymbol{\eta}_Y)\pi^a(1-\pi)^{(1-a)} f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\eta}_{\boldsymbol{X}}) : \boldsymbol{\eta} = (\boldsymbol{\eta}_Y, \boldsymbol{\eta}_{\boldsymbol{X}}) \Big\} \quad (6.8)$$

with $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_{\boldsymbol{X}}$ infinite dimensional nuisance parameters. We use semiparametric theory to derive the class of all consistent and asymptotically normal estimators for ν_0 under model $\mathcal{M}(\pi)$, which is outlined in Chapter 2. We aim to identify the class of all asymptotically linear estimators $\hat{\nu}_n$ for ν_0, thus all estimators that can be written as $n^{1/2}(\hat{\nu}_n - \nu_0) = \sum_{i=1}^n n^{-1/2}\phi(Y_i, A_i, \boldsymbol{X}_i; \nu_0) + o_p(1)$, where ϕ is the influence function that satisfies $E(\phi) = 0$ and $E(\phi^2) < \infty$ and $o_p(1)$ is a term that converges to zero in probability. By Slutsky's theorem it then follows that the asymptotic distribution of $n^{1/2}(\hat{\nu}_n - \nu_0)$ is given by $N\{0, E(\phi^2)\}$. The aim is thus to identify all influence functions for ν_0. Given one specific influence function ϕ_0 for ν_0, the space of all influence functions for ν_0 is given by the set $\phi_0 + \mathcal{T}(\pi)^\perp$ (see Theorem 2.11) with $\mathcal{T}(\pi)^\perp$ the orthogonal complement of the tangent space $\mathcal{T}(\pi)$ of the model $\mathcal{M}(\pi)$. The tangent space can be written as the direct sum $\mathcal{T}_Y(\pi) \oplus \mathcal{T}_{\boldsymbol{X}}(\pi)$ with $\mathcal{T}_Y(\pi)$ the tangent space corresponding to the conditional density function $f_{Y|A,\boldsymbol{X}}(y|a,\boldsymbol{x};\boldsymbol{\eta}_Y)$ and $\mathcal{T}_{\boldsymbol{X}}(\pi)$ the tangent space corresponding to the marginal density function $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\eta}_{\boldsymbol{X}})$. It follows from Theorem 2.12 that $\mathcal{T}_Y(\pi) = \big\{\alpha_Y(Y,A,\boldsymbol{X}) : E\{\alpha_Y(Y,A,\boldsymbol{X})|A,\boldsymbol{X}\} = 0\big\}$ and $\mathcal{T}_{\boldsymbol{X}}(\pi) = \big\{\alpha_{\boldsymbol{X}}(\boldsymbol{X}) : E\{\alpha_{\boldsymbol{X}}(\boldsymbol{X})\} = 0\big\}$ with $\alpha_Y(Y,A,\boldsymbol{X})$ and $\alpha_{\boldsymbol{X}}(\boldsymbol{X})$ square-integrable. Note that $\mathcal{T}_Y(\pi) \perp \mathcal{T}_{\boldsymbol{X}}(\pi)$. From this, it follows that $\mathcal{T}(\pi)^\perp = \big\{\alpha(A,\boldsymbol{X}) : E\{\alpha(A,\boldsymbol{X})|\boldsymbol{X}\} = 0\big\}$ with $\alpha(A,\boldsymbol{X})$ square-integrable. Because A is binary, $\alpha(A,\boldsymbol{X}) = A\alpha(1,\boldsymbol{X}) + (1-A)\alpha(0,\boldsymbol{X})$ and from $E\{\alpha(A,\boldsymbol{X})|\boldsymbol{X}\} = 0$, it follows that this space can be equivalently written as $\mathcal{T}(\pi)^\perp = \big\{(A-\pi)\tilde{\alpha}(\boldsymbol{X}) : \tilde{\alpha}(\boldsymbol{X})$ arbitrary square-integrable function of $\boldsymbol{X}\big\}$. To conclude, the set of all influence functions for ν_0 is given by $\big\{\phi_0 + (A-\pi)\tilde{\alpha}(\boldsymbol{X})\big\}$ with ϕ_0 an arbitrary influence function for ν_0. For instance, one may let ϕ_0 be the influence function of the initial estimator $\hat{\nu}_{n,0} = \{n(n-1)\}^{-1}\sum_{i=1}^n \sum_{j\neq i}(1 - A_i)A_j I_{ij}/\{\pi(1-\pi)\}$ with $I_{ij} = I(Y_i \preceq Y_j)$. To derive ϕ_0, we consider the Hájek

projection (Hájek 1970; van der Vaart 1998) of $\hat{\nu}_{n,0}$, which is given by

$$\begin{aligned}\tilde{\nu}_{n,0} &= \sum_{i=1}^{n} E(\hat{\nu}_{n,0}|Y_i,A_i,\boldsymbol{X}_i) - \nu_0 \\ &= n^{-1}\sum_{i=1}^{n}\frac{1-A_i}{1-\pi}a_1(Y_i) - \nu_0 + \frac{A_i}{\pi}a_2(Y_i) - \nu_0, \end{aligned} \tag{6.9}$$

with $a_1(Y_i) = E\{(A_j/\pi)I_{ij}|Y_i\}$ and $a_2(Y_i) = E\{(1-A_j)/(1-\pi)I_{ji}|Y_i\}$. Under regularity conditions ($E\left([(1-A_i)A_jI_{ij}/\{\pi(1-\pi)\}]^2\right) < \infty$), it follows from Theorem 12.3 in van der Vaart (1998) that $n^{1/2}(\hat{\nu}_{n,0} - \nu_0 - \tilde{\nu}_{n,0}) \xrightarrow{P} 0$. Consequently, we get

$$n^{1/2}(\hat{\nu}_{n,0} - \nu_0) = n^{1/2}\tilde{\nu}_{n,0} + o_p(1) = n^{-1/2}\sum_{i=1}^{n}\phi_0(Y_i,A_i;\nu_0) + o_p(1), \tag{6.10}$$

with $\phi_0(Y_i,A_i;\nu_0) = \{(1-A_i)/(1-\pi)\}a_1(Y_i) - \nu_0 + (A_i/\pi)a_2(Y_i) - \nu_0$. We conclude that every consistent and regular asymptotically linear estimator of ν_0 has an influence function in the class

$$\left\{\phi(Y,A,\boldsymbol{X};\nu_0) = \frac{1-A}{1-\pi}a_1(Y) - \nu_0 + \frac{A}{\pi}a_2(Y) - \nu_0 + (A-\pi)\tilde{\alpha}(\boldsymbol{X}) : \right.$$
$$\left. \tilde{\alpha}(\boldsymbol{X}) \text{ arbitrary square-integrable function of } \boldsymbol{X}\right\}. \tag{6.11}$$

For an arbitrary square-integrable function $H(\boldsymbol{X}_i,\boldsymbol{X}_j)$, the estimator

$$\hat{\nu}_n(H;\pi) = \hat{\nu}_{n,0} + \{n(n-1)\}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\left\{1 - \frac{(1-A_i)A_j}{(1-\pi)\pi}\right\}H(\boldsymbol{X}_i,\boldsymbol{X}_j) \tag{6.12}$$

has influence function in the set (6.11) with

$$\tilde{\alpha}(\boldsymbol{X}_i) = \left[\frac{E\{H(\boldsymbol{X}_i,\boldsymbol{X}_j)|\boldsymbol{X}_i\}}{1-\pi} - \frac{E\{H(\boldsymbol{X}_j,\boldsymbol{X}_i)|\boldsymbol{X}_i\}}{\pi}\right], \tag{6.13}$$

which can be shown by taking its Hájek projection.

Class of all consistent and asymptotically normal estimators for ν_0 under model $\mathcal{M}_{\text{indep}}$

To simplify the above derivation, we used the true randomization probability π instead of the observed proportion $\hat{\pi}_n = N_1/n$. Nevertheless, more efficient estimators can be obtained by estimating π (Robins et al. 1992; Rotnitzky et al. 2010). In this paragraph, we therefore consider the model $\mathcal{M}_{\text{indep}}$, which only assumes $A \perp\!\!\!\perp \boldsymbol{X}$ with π unknown. This semiparametric model can be formalized as the set of all joint laws

$$\begin{aligned}\mathcal{M}_{\text{indep}} = \Big\{ & f_{Y,A,\boldsymbol{X}}(y,a,\boldsymbol{x};\pi,\boldsymbol{\eta}) \\ & = f_{Y|A,\boldsymbol{X}}(y|a,\boldsymbol{x};\boldsymbol{\eta}_Y)\pi^a(1-\pi)^{(1-a)} f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\eta}_{\boldsymbol{X}}) : \pi, \boldsymbol{\eta} = (\boldsymbol{\eta}_Y, \boldsymbol{\eta}_{\boldsymbol{X}}) \Big\}\end{aligned} \tag{6.14}$$

with $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_{\boldsymbol{X}}$ infinite dimensional nuisance parameters but with the subtle difference that $0 < \pi < 1$ is now also an unknown parameter. The tangent space of model $\mathcal{M}_{\text{indep}}$ is $\mathcal{T} = \mathcal{T}(\pi) \oplus \mathcal{T}_A$ where $\mathcal{T}_A = \{\alpha_\pi S_\pi : \alpha_\pi \in \mathbb{R}\}$ with $S_\pi = (A-\pi)/\{\pi(1-\pi)\}$, the score function for π. Because $\mathcal{T}_A \subset \mathcal{T}(\pi)^\perp$, $\mathcal{T}_A \perp \mathcal{T}(\pi)$. An influence function for ν_0 under model $\mathcal{M}_{\text{indep}}$ can hence be obtained as the residual after projecting ϕ from (6.11) onto $\mathcal{T}_A$, $\phi_{\text{est}} = \phi - \Pi(\phi|\mathcal{T}_A) = \phi - E(\phi S_\pi)E^{-1}(S_\pi^2)S_\pi$ with $E(\phi S_\pi) = E\{a_2(Y)|A=1\}/\pi - E\{a_1(Y)|A=0\}/(1-\pi) + E\{\tilde{\alpha}(\boldsymbol{X})\}$ and $E^{-1}(S_\pi^2) = \pi(1-\pi)$. The influence function is thus given by $\phi_{\text{est}} = \phi_0 + (A-\pi)\tilde{\alpha}_{\text{est}}(\boldsymbol{X})$ with $\tilde{\alpha}_{\text{est}}(\boldsymbol{X}) = \tilde{\alpha}(\boldsymbol{X}) - E(\phi S_\pi) = \tilde{\alpha}(\boldsymbol{X}) - E\{\tilde{\alpha}(\boldsymbol{X})\} + [E\{a_1(Y)|A=0\}/(1-\pi) - E\{a_2(Y)|A=1\}/\pi]$. Because ϕ_{est} is a projection, $E(\phi_{\text{est}}^2) \leq E(\phi^2)$ leading to a smaller asymptotic variance. The estimator

$$\hat{\nu}_n(H) = \frac{U}{N_0 N_1} + \sum_{i=1}^{n}\sum_{j\neq i}\left\{\frac{1}{n(n-1)} - \frac{(1-A_i)A_j}{N_0 N_1}\right\} H(\boldsymbol{X}_i, \boldsymbol{X}_j)$$

equals $\hat{\nu}_n(H;\hat{\pi}_n)$ given in (6.12). From the previous results, its influence function under $\mathcal{M}_{\text{indep}}$ equals ϕ_{est} with $\tilde{\alpha}_{\text{est}}(\boldsymbol{X}) = \tilde{\alpha}(\boldsymbol{X}) - E(\phi S_\pi)$ and $\tilde{\alpha}(\boldsymbol{X})$ given in (6.13).

We thus have proven the following Theorem:

Theorem 6.2. *Under model $\mathcal{M}_{indep}$, all consistent and asymptotically normal estimators for the MPI are asymptotically equivalent to an estimator from the class*

$$\hat{v}_n(H) = \frac{U}{N_0 N_1} + \sum_{i=1}^{n} \sum_{j \neq i} \left\{ \frac{1}{n(n-1)} - \frac{(1-A_i)A_j}{N_0 N_1} \right\} H(\boldsymbol{X}_i, \boldsymbol{X}_j), \qquad (6.15)$$

where $H(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is an arbitrary square-integrable function of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$.

This class of estimators **augments** the standardized Mann-Whitney U statistic U/N_0N_1 by incorporating covariate adjustment in a similar vein as for marginal additive treatment effects (Tsiatis et al. 2008; Zhang et al. 2008). It is not difficult to see that every choice of $H(\boldsymbol{X}_i, \boldsymbol{X}_j)$ yields a consistent estimator $\hat{v}_n(H)$ for the MPI v_0. This is because $A \perp\!\!\!\perp \boldsymbol{X}$ as implied by randomization, and hence the expectation of the augmentation term equals

$$E\left\{1 - (1-A_i)A_j / [N_0 N_1 / \{n(n-1)\}]\right\} E\{H(\boldsymbol{X}_i, \boldsymbol{X}_j)\},$$

which converges to zero for every choice of $H(\boldsymbol{X}_i, \boldsymbol{X}_j)$. From (6.15) it is now clear that the function $H(\boldsymbol{X}_i, \boldsymbol{X}_j)$ dictates the nature of covariate adjustment for estimators within this class, with a constant corresponding to no adjustment at all. In particular, for $H(\boldsymbol{X}_i, \boldsymbol{X}_j) \equiv c$ for some constant c, (6.15) reduces to the standardized Mann-Whitney U statistic. Different choices of $H(\boldsymbol{X}_i, \boldsymbol{X}_j)$ thus lead to estimators $\hat{v}_n(H)$ with different asymptotic behavior.

Efficient consistent and asymptotically normal estimator $\hat{v}_n(H_{\text{eff}})$

The Theorem below shows that efficient estimation of the MPI under model $\mathcal{M}_{\text{indep}}$ relies on the CPI $P(Y \preceq Y^* | A = 0, A^* = 1, \boldsymbol{X}, \boldsymbol{X}^*)$, just like efficient estimation of the additive treatment effect demands the conditional expectations $E(Y|A = k, \boldsymbol{X})$, $k = 0, 1$ (Tsiatis et al. 2008).

Theorem 6.3. *The choice*

$$H_{eff}(\boldsymbol{X}_i, \boldsymbol{X}_j) = P(Y_i \preceq Y_j | A_i = 0, A_j = 1, \boldsymbol{X}_i, \boldsymbol{X}_j) \qquad (6.16)$$

yields $\hat{\nu}_n(H_{eff})$ to achieve the semiparametric variance bound for the MPI under model $\mathcal{M}_{indep}$.

Proof. The asymptotically efficient estimator of ν_0 under model $\mathcal{M}(\pi)$ has the influence function in the class (6.11) with the smallest variance. From Theorem 2.11, it follows that the efficient influence function is given by the projection of any influence function, e.g., ϕ_0, onto the model tangent space: $\phi_{\text{eff}} = \Pi\{\phi_0|\mathcal{T}(\pi)\} = \phi_0 - E(\phi_0|A,\boldsymbol{X}) + E(\phi_0|\boldsymbol{X}) = \phi_0 + (A-\pi)\tilde{\alpha}_{\text{eff}}(\boldsymbol{X})$ with $\tilde{\alpha}_{\text{eff}}(\boldsymbol{X}) = E\{a_1(Y)|A = 0,\boldsymbol{X}\}/(1-\pi) - E\{a_2(Y)|A = 1,\boldsymbol{X}\}/\pi$. Because

$$E\{a_1(Y_i)|A_i = 0,\boldsymbol{X}_i\} = E\{P(Y_i \preceq Y_j|A_i = 0,A_j = 1,\boldsymbol{X}_i,\boldsymbol{X}_j)|\boldsymbol{X}_i\},$$
$$E\{a_2(Y_i)|A_i = 1,\boldsymbol{X}_i\} = E\{P(Y_j \preceq Y_i|A_j = 0,A_i = 1,\boldsymbol{X}_j,\boldsymbol{X}_i)|\boldsymbol{X}_i\},$$

$H_{\text{eff}}(\boldsymbol{X}_i,\boldsymbol{X}_j) = P(Y_i \preceq Y_j|A_i = 0,A_j = 1,\boldsymbol{X}_i,\boldsymbol{X}_j)$. It follows that $\hat{\nu}_n(H_{\text{eff}};\pi)$ is the most efficient estimator of ν under model $\mathcal{M}(\pi)$. Because $\phi_{\text{eff}} \in \mathcal{T}(\pi)$ and $\mathcal{T}_A \subset \mathcal{T}(\pi)^{\perp}$, $\phi_{\text{eff,est}} \equiv \phi_{\text{eff}}$. This means that the efficient influence function for ν_0 under model $\mathcal{M}_{\text{indep}}$ is the same as under model $\mathcal{M}(\pi)$ where the randomization probabilities are known. Consequently, the estimator $\hat{\nu}_n(H_{\text{eff}}) \equiv \hat{\nu}_n(H_{\text{eff}};\hat{\pi}_n)$ has the same influence function as $\hat{\nu}_n(H_{\text{eff}};\pi)$. We conclude $\hat{\nu}_n(H_{\text{eff}})$ is an asymptotically efficient estimator of ν_0 under model $\mathcal{M}_{\text{indep}}$. □

6.5.2 Practical implementation: a locally efficient and adaptive estimation strategy

Since the conditional probabilistic index is not known in practice, it must be modeled. The efficiency result of the previous section is then local: attained under correct specification of a model for $P(Y_i \preceq Y_j|A_i = 0,A_j = 1,\boldsymbol{X}_i,\boldsymbol{X}_j)$. We here propose a locally efficient adaptive estimation strategy for the MPI ν_0 and explain the connection with the imputation estimator $\hat{\nu}_{n,\text{IMP}}$ from Section 6.4. In Appendix 6.D.1, we provide an R-function that implements the methods below.

Locally efficient and adaptive strategy

Step 1. Postulate a PIM $m(A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})$ as a working model for the conditional probabilistic index $P(Y_i \preceq Y_j|A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j)$ such as models (6.5) or (6.6). Let $\hat{\boldsymbol{\tau}}_n$ be the estimator for the regression coefficients $\boldsymbol{\tau}$ obtained by solving the estimating equations (6.22). For each pair (i,j) with $i \neq j$, $i,j = 1,\ldots,n$, obtain predictions $m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)$.

Step 2. Given the predictions from the working model, the locally efficient estimator for the MPI can be calculated as

$$\hat{v}_{n,\text{adap}} = \frac{U}{N_0N_1} + \sum_{i=1}^{n}\sum_{j\neq i}\left[\left\{\frac{1}{n(n-1)} - \frac{(1-A_i)A_j}{N_0N_1}\right\} \times m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)\right], \quad (6.17)$$

so that $\hat{v}_{n,\text{adap}} \equiv \hat{v}_n(\hat{m}_n)$, $\hat{m}_n \equiv m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)$. For instance, using (6.5) as a working model, (6.17) becomes

$$\hat{v}_{n,\text{adap}} = \frac{U}{N_0N_1} + \sum_{i=1}^{n}\sum_{j\neq i}\left[\left\{\frac{1}{n(n-1)} - \frac{(1-A_i)A_j}{N_0N_1}\right\} \times \text{expit}\{\hat{\tau}_{n,A} + \hat{\boldsymbol{\tau}}_{n,\boldsymbol{X}}^T(\boldsymbol{X}_j - \boldsymbol{X}_i)\}\right].$$

Effect of parameter estimation in the postulated model for $P(Y_i \preceq Y_j|A_i = 0,A_j = 1,\boldsymbol{X}_i,\boldsymbol{X}_j)$

If the working model $m(A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})$ is correctly specified, then under standard regularity conditions, the estimator $\hat{v}_{n,\text{adap}}$ is asymptotically equivalent to the unfeasible estimator $\hat{v}_n(H_{\text{eff}})$ and thus asymptotically efficient. Substitution of estimators for the regression coefficients in the PIM thus leads to an estimator for v_0 with the same asymptotic variance as if the CPI $H_{\text{eff}}(\boldsymbol{X}_i,\boldsymbol{X}_j)$ were known, making it adaptive. If the working model $m(A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})$ is misspecified, then $\hat{v}_{n,\text{adap}}$ remains consistent. However, since it is based on a different function $H(\boldsymbol{X}_i,\boldsymbol{X}_j)$, namely $m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau}^*)$ with $\boldsymbol{\tau}^*$ the probability limit of some estimator for $\boldsymbol{\tau}$, it

is no longer efficient. Nevertheless, substitution of estimators for the regression coefficients in the misspecified PIM also leads an estimator for ν_0 with the same asymptotic variance as if $\boldsymbol{\tau}^*$ were known.

Theorem 6.4. *The asymptotic behavior of the estimator $\hat{\nu}_{n,adap} \equiv \hat{\nu}_n(\hat{m}_n)$, $\hat{m}_n \equiv m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)$ is the same as that of $\hat{\nu}_{n,adap} \equiv \hat{\nu}_n(m^*)$, $m^* \equiv m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau}^*)$, with $\boldsymbol{\tau}^* = plim(\hat{\boldsymbol{\tau}}_n)$. In particular, under correct specification of the working PIM, $\hat{\nu}_{n,adap}$ is asymptotically equivalent to $\hat{\nu}_n(H_{eff})$.*

Proof. Under suitable regularity conditions, we have that $\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^* = O_p(n^{-1/2})$ ($n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*)$ is bounded in probability) with $\boldsymbol{\tau}^*$ the probability limit of $\hat{\boldsymbol{\tau}}_n$. When the working PIM is correctly specified, $\boldsymbol{\tau}^*$ is the value satisfying $P(Y_i \preceq Y_j|A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j) = m(A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau}^*)$. Assuming sufficient regularity conditions, we have

$$\begin{aligned}
n^{1/2}(\hat{\nu}_{n,\text{adap}} - \nu_0) =& n^{-1/2}\sum_{i=1}^{n}\phi_{\text{est}}(Y_i,A_i,\boldsymbol{X}_i;\nu_0,\boldsymbol{\tau}^*) + o_p(1) \\
&+ \left\{ n^{-1}\sum_{i=1}^{n}\phi_{\text{est},\boldsymbol{\tau}}(Y_i,A_i,\boldsymbol{X}_i;\nu_0,\tilde{\boldsymbol{\tau}}_n)\right\} n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*) \\
=& n^{-1/2}\sum_{i=1}^{n}\{\phi_0(Y_i,A_i;\nu_0) + (A_i-\pi)\tilde{\alpha}_{\text{est}}(\boldsymbol{X}_i;\boldsymbol{\tau}^*)\} + o_p(1) \\
&+ \left\{ n^{-1}\sum_{i=1}^{n}(A_i-\pi)\tilde{\alpha}_{\text{est},\boldsymbol{\tau}}(\boldsymbol{X}_i;\tilde{\boldsymbol{\tau}}_n)\right\} n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*)
\end{aligned}$$

where $\tilde{\boldsymbol{\tau}}_n$ is an intermediate value on the line segment connecting $\hat{\boldsymbol{\tau}}_n$ and $\boldsymbol{\tau}^*$,

$$\begin{aligned}
\phi_{\text{est},\boldsymbol{\tau}}(Y_i,A_i,\boldsymbol{X}_i;\nu_0,\tilde{\boldsymbol{\tau}}_n) &= \partial\phi_{\text{est}}(Y_i,A_i,\boldsymbol{X}_i;\nu_0,\boldsymbol{\tau})/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\tilde{\boldsymbol{\tau}}_n}, \\
\tilde{\alpha}_{\text{est}}(\boldsymbol{X}_i;\boldsymbol{\tau}^*) &= \tilde{\alpha}(\boldsymbol{X}_i;\boldsymbol{\tau}^*) - E\{\tilde{\alpha}(\boldsymbol{X}_i;\boldsymbol{\tau}^*)\} \\
&\quad + \frac{E\{a_1(Y_i)|A_i=0\}}{1-\pi} - \frac{E\{a_2(Y_i)|A_i=1\}}{\pi}, \\
\tilde{\alpha}(\boldsymbol{X}_i;\boldsymbol{\tau}^*) &= \frac{E\{m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau}^*)|\boldsymbol{X}_i\}}{1-\pi} - \frac{E\{m(0,1,\boldsymbol{X}_j,\boldsymbol{X}_i;\boldsymbol{\tau}^*)|\boldsymbol{X}_i\}}{\pi}, \\
\tilde{\alpha}_{\text{est},\boldsymbol{\tau}}(\boldsymbol{X}_i;\tilde{\boldsymbol{\tau}}_n) &= \partial\tilde{\alpha}_{\text{est}}(\boldsymbol{X}_i;\boldsymbol{\tau})/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\tilde{\boldsymbol{\tau}}_n}.
\end{aligned}$$

The term $n^{-1}\sum_{i=1}^{n}(A_i-\pi)\tilde{\alpha}_{\text{est},\boldsymbol{\tau}}(\boldsymbol{X}_i;\tilde{\boldsymbol{\tau}}_n)$ converges in probability to zero because $E\{(A-\pi)\tilde{\alpha}_{\text{est},\boldsymbol{\tau}}(\boldsymbol{X};\boldsymbol{\tau}^*)\} = E(A-\pi)E\{\tilde{\alpha}_{\text{est},\boldsymbol{\tau}}(\boldsymbol{X};\boldsymbol{\tau}^*)\} = \mathbf{0}$ since $A \perp\!\!\!\perp \boldsymbol{X}$. Because $\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^* = O_p(n^{-1/2})$, this shows that $\hat{v}_{n,\text{adap}} = \hat{v}_n(\hat{m}_n)$ and $\hat{v}_n(m^*)$ have the same limit distribution, with $\hat{m}_n \equiv m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)$ and $m^* \equiv m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau}^*)$. The asymptotic behavior of $\hat{v}_{n,\text{adap}}$ is thus the same as if $\boldsymbol{\tau}^*$ were known, regardless of correct specification of the working PIM. □

Sandwich estimator for the asymptotic variance of $\hat{v}_{n,\text{adap}}$ and inference

The estimator $\hat{v}_{n,\text{adap}}$ has influence function $\phi_{\text{est}}(Y,A,\boldsymbol{X};v_0,\boldsymbol{\tau}^*)$. Its asymptotic variance is then given by $E\{\phi^2_{\text{est}}(Y,A,\boldsymbol{X};v_0,\boldsymbol{\tau}^*)\}$. In moderate to large samples, the variance of $\hat{v}_{n,\text{adap}}$ can thus be estimated as one over n times the estimated variance of the influence function:

$$\begin{aligned}\widehat{\text{var}}_n(\hat{v}_{n,\text{adap}}) &= n^{-2}\sum_{i=1}^{n}\hat{\phi}^2_{n,\text{est}}(Y_i,A_i,\boldsymbol{X}_i;\hat{v}_{n,\text{adap}},\hat{\boldsymbol{\tau}}_n)\\ &= n^{-2}\sum_{i=1}^{n}\left\{\hat{\phi}_{n,0}(Y_i,A_i;\hat{v}_{n,\text{adap}}) + (A_i-\hat{\pi}_n)\hat{\tilde{\alpha}}_{n,\text{est}}(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)\right\}^2 \qquad (6.18)\end{aligned}$$

with

$$\begin{aligned}\hat{\phi}_{n,0}(Y_i,A_i;\hat{v}_{n,\text{adap}}) &= \frac{1-A_i}{1-\hat{\pi}_n}\hat{a}_{n,1}(Y_i) - \hat{v}_{n,\text{adap}} + \frac{A_i}{\hat{\pi}_n}\hat{a}_{n,2}(Y_i) - \hat{v}_{n,\text{adap}},\\ \hat{\tilde{\alpha}}_{n,\text{est}}(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n) &= \hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n) - \hat{E}_n\{\hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)\}\\ &\qquad + \frac{\hat{E}_n\{\hat{a}_{n,1}(Y_i)|A_i=0\}}{1-\hat{\pi}_n} - \frac{\hat{E}_n\{\hat{a}_{n,2}(Y_i)|A_i=1\}}{\hat{\pi}_n},\\ \hat{a}_{n,1}(Y_i) &= (n-1)^{-1}\sum_{j\neq i}A_j I_{ij}/\hat{\pi}_n,\\ \hat{a}_{n,2}(Y_i) &= (n-1)^{-1}\sum_{j\neq i}(1-A_j)I_{ji}/(1-\hat{\pi}_n),\\ \hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\tau}_n) &= (n-1)^{-1}\sum_{j\neq i}\{\hat{m}_{n,ij}/(1-\hat{\pi}_n) - \hat{m}_{n,ji}/\hat{\pi}_n\},\\ \hat{m}_{n,ij} &= m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n),\\ \hat{E}_n\{\hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)\} &= n^{-1}\sum_{i=1}^{n}\hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n),\end{aligned}$$

$$\hat{E}_n\{\hat{a}_{n,1}(Y_i)|A_i=0\} = n^{-1}\sum_{i=1}^{n}\frac{1-A_i}{1-\hat{\pi}_n}\hat{a}_{n,1}(Y_i) = N_0^{-1}\sum_{i=1}^{n}(1-A_i)\hat{a}_{n,1}(Y_i),$$
$$\hat{E}_n\{\hat{a}_{n,2}(Y_i)|A_i=1\} = n^{-1}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}_n}\hat{a}_{n,2}(Y_i) = N_1^{-1}\sum_{i=1}^{n}A_i\hat{a}_{n,2}(Y_i).$$

In the calculation of the sandwich estimator, $\hat{\tilde{\alpha}}_{n,\text{est}}(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)$ is used instead of $\hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)$ to acknowledge the estimation of the randomization probabilities. This prevents conservative inference for the target parameter under misspecification of the PIM. A Wald $(1-\alpha)100\%$ confidence interval for ν_0 may then be obtained as $\hat{\nu}_{n,\text{adap}} \pm z_{\alpha/2}\{\widehat{\text{var}}_n(\hat{\nu}_{n,\text{adap}})\}^{1/2}$ with $z_{\alpha/2}$ such that $\Phi(z_{\alpha/2}) = 1-\alpha/2$ and $\Phi(\cdot)$ the cumulative distribution function of a standard normal random variable. The performance of the estimator (6.17) and the corresponding sandwich estimator will be illustrated in simulation studies (Section 6.8). To test the null hypothesis of no treatment effect, one can use a Wald test by relying on the asymptotic normality of the test statistic $Z = (\hat{\nu}_{n,\text{adap}} - 0.5)/\{\widehat{\text{var}}_n(\hat{\nu}_{n,\text{adap}})\}^{1/2}$ and reject if $|Z| > z_{\alpha/2}$. Asymptotically, the Wald test based on this locally efficient estimator will be more powerful than the classical Mann-Whitney U test (that is, choosing $H(\boldsymbol{X}_i,\boldsymbol{X}_j)$ to be a constant). However, because the Mann-Whitney U test is often indicated in small samples where the sandwich estimator may not well approximate the true sampling variability, especially when covariate selection is applied, we propose a permutation test in Section 6.6 to test for the absence of a treatment effect.

Equivalence $\hat{\nu}_{n,\textbf{adap}}$ and $\hat{\nu}_{n,\textbf{IMP}}$

In Section 6.4, we demonstrated how an estimator for the MPI can be obtained by standardization of the CPI. This standardized CPI (6.7) equals $\hat{\nu}_{n,\text{adap}}$ if the PIM includes a main effect $\tau_A(A_j - A_i)$ for the treatment and a logit link function is used and estimating equations (6.22) are used to estimate the regression parameters $\boldsymbol{\tau}$ indexing the PIM. Indeed, in that case it follows from the estimating equations (6.22) that

$$\sum_{i=1}^{n}\sum_{j\neq i}(A_j - A_i)[I_{ij} - m(A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)] = 0.$$

Rewriting $(A_j - A_i)$ as $A_j(1-A_i) - A_i(1-A_j)$ gives

$$0 = \sum_{i=1}^{n} \sum_{j \neq i} A_j(1-A_i)\{I_{ij} - m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)\} + \sum_{i=1}^{n} \sum_{j \neq i} A_i(1-A_j)\{I_{ji} - m(0,1,\boldsymbol{X}_j,\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)\}$$

because $I_{ij} = 1 - I_{ji}$ and $m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n) = 1 - m(1,0,\boldsymbol{X}_j,\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)$. Interchanging i and j in the second summation of the latter equation gives

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \sum_{j \neq i} A_j(1-A_i)[I_{ij} - m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n)] \\ &= U - \sum_{i=1}^{n} \sum_{j \neq i} A_j(1-A_i) m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n). \end{aligned}$$

6

Because $\hat{v}_{n,\text{adap}}$ can be written as

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n) + \frac{1}{N_0 N_1} \left\{ U - \sum_{i=1}^{n} \sum_{j \neq i} A_j(1-A_i) m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n) \right\},$$

we conclude that under these conditions $\hat{v}_{n,\text{adap}} = \hat{v}_{n,\text{IMP}}$, the mean of all predicted pairwise comparisons. This estimator has the further advantage that it is guaranteed to fall in the parameter space $(0,1)$, which is not guaranteed for all estimators of the form (6.17). Additionally, because $\hat{v}_{n,\text{IMP}}$ does not involve the randomization probabilities, one does not need to correct for their estimation, simplifying the calculation of the sandwich estimator. In this case, $\hat{\tilde{\alpha}}_{n,\text{est}}(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n) = \hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n) + o_p(1)$ (see page 194):

$$\begin{aligned} &\hat{\tilde{\alpha}}_{n,\text{est}}(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n) - \hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n) \\ &\quad = -\hat{E}_n\{\hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)\} + \hat{E}_n\{\hat{a}_{n,1}(Y_i)|A_i=0\}/(1-\hat{\pi}_n) - \hat{E}_n\{\hat{a}_{n,2}(Y_i)|A_i=1\}/\hat{\pi}_n \\ &\quad = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \left[\frac{1}{1-\hat{\pi}_n} \left\{ \frac{(1-A_i)A_j I_{ij}}{(1-\hat{\pi}_n)\hat{\pi}_n} - m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n) \right\} \right. \end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{1}{\hat{\pi}_n}\left\{\frac{(1-A_j)A_iI_{ji}}{(1-\hat{\pi}_n)\hat{\pi}_n} - m(0,1,\boldsymbol{X}_j,\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)\right\}\Bigg] \\
&= \frac{2\hat{\pi}_n - 1}{\hat{\pi}_n(1-\hat{\pi}_n)}\left\{\frac{U}{N_0N_1} - \hat{v}_{n,\text{IMP}} + o_p(1)\right\},
\end{aligned}
$$

which is $o_p(1)$ because $U/(N_0N_1) = v_0 + o_p(1)$ and $\hat{v}_{n,\text{IMP}} = v_0 + o_p(1)$, where the latter equality also holds under misspecification of the PIM given the PIM is defined using a logit link, includes a main effect $\tau_A(A_j - A_i)$ for the treatment and is estimated via (6.22). We can conclude that $\hat{\tilde{\alpha}}_n(\boldsymbol{X}_i;\hat{\tau}_n)$ can be used instead of $\hat{\tilde{\alpha}}_{n,\text{est}}(\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}_n)$.

6.5.3 Connection to *the improved hypothesis tests of Zhang et al. (2008)*

The Wald test based on the test statistic $Z = (\hat{v}_{n,\text{adap}} - 0.5)/\{\widehat{\text{var}}_n(\hat{v}_{n,\text{adap}})\}^{1/2}$ is also related to the improved Kruskal-Wallis test in Zhang et al. (2008), Section 5, which reduces to the Mann-Whitney test for two groups. In this manuscript, an inefficient estimator for the MPI is augmented via a working model for the CPI using PIMs. In contrast, in Zhang et al. (2008), augmentation of an asymptotically linear test statistic is performed (which is asymptotically equivalent to the Mann-Whitney U test statistic) using two different linear working models for the survival function of the outcomes within both treatment groups. Specifically, the procedure of Zhang et al. (2008) proceeds by augmenting $\ell(Y,A) = (A-\pi)\{S(Y) - 0.5\}$, with $S(y) = 1 - P(Y \leq y)$ the survival function of Y, to incorporate covariate information. The optimal augmentation equals $\ell^*(Y,\boldsymbol{X},A) = \ell(Y,A) + (A-\pi)[E\{\ell(Y,0)|\boldsymbol{X},A = 0\} - E\{\ell(Y,1)|\boldsymbol{X},A = 1\}]$. Based on ℓ^*, a test statistic $\widehat{T}_n^*$ is constructed,

$$
\widehat{T}_n^* = \left\{n^{-1/2}\sum_{i=1}^n \hat{\ell}_n^*(Y_i,\boldsymbol{X}_i,A_i)\right\}^2 \times \left\{n^{-1}\sum_{i=1}^n \hat{\ell}_n^*(Y_i,\boldsymbol{X}_i,A_i)^2\right\}^{-1},
$$

with $\hat{\ell}_n^*(Y_i,\boldsymbol{X}_i,A_i) = \hat{\ell}_n(Y_i,A_i) + (A_i - \hat{\pi}_n)\{q_0(\boldsymbol{X}_i;\hat{\boldsymbol{\zeta}}_{n,0}) - q_1(X_i;\hat{\boldsymbol{\zeta}}_{n,1})\}$, $\hat{\ell}_n(Y_i,A_i) = (A_i - \hat{\pi}_n)\{\hat{S}_n(Y_i) - 0.5\}$, $\hat{\pi}_n = n^{-1}\sum_{i=1}^n A_i$, $\hat{S}_n(y) = n^{-1}\sum_{i=1}^n I(Y_i \geq y)$ and with

working models

$$E\{\hat{\ell}_n(Y_i,0)|\boldsymbol{X}_i,A_i=0;\boldsymbol{\zeta}_0\}=q_0(\boldsymbol{X}_i;\boldsymbol{\zeta}_0)=\zeta_{00}+\boldsymbol{\zeta}_{01}^T\boldsymbol{X}_i,$$
$$E\{\hat{\ell}_n(Y_i,1)|\boldsymbol{X}_i,A_i=1;\boldsymbol{\zeta}_1\}=q_1(\boldsymbol{X}_i;\boldsymbol{\zeta}_1)=\zeta_{10}+\boldsymbol{\zeta}_{11}^T\boldsymbol{X}_i.$$

The latter approach has the disadvantages that a linear model for the survival function does not respect the boundaries of the quantity it is estimating, is less interpretable than a working model for the CPI, that such a model is more likely to be misspecified and that it does not deliver an estimator for the MPI. The performance of both tests will be illustrated in simulation studies (Section 6.8). In Appendix 6.C, we elaborate on the mathematical details connecting both methodologies.

6.6 Randomization Inference: Augmentation of the Mann-Whitney U Test

In this section, we propose a permutation test based on the locally efficient estimator $\hat{\nu}_{n,\text{adap}}$ (6.17) for the MPI. The motivation and rationale behind this is that we anticipate a permutation test based on a locally efficient estimator for the MPI ($\hat{\nu}_{n,\text{adap}}$) under model $\mathcal{M}_{\text{indep}}$ to have higher power than a test based on an inefficient estimator for the MPI (U/N_0N_1) under model $\mathcal{M}_{\text{indep}}$, even though the increase in efficiency and power is only guaranteed asymptotically. This will be illustrated in an application to an HIV clinical trial (Section 6.7) and in simulation studies (Section 6.8). For a gentle introduction and overview of randomization and permutation methods, we refer to Ernst (2004). For a more thorough discussion, we refer to Lehmann and Romano (2005), chap. 15 or Boos and Stefanski (2013), chap. 12.

The null hypothesis that the distribution of the observables $(Y,\boldsymbol{X})$ is the same in both treatment groups,

$$H_0: A \perp\!\!\!\perp (Y,\boldsymbol{X}), \tag{6.19}$$

states that treatment is not effective. In Section 6.6.1, we give the theoretical construction and justification of the proposed permutation test of the null hypothesis

(6.19) and in Section 6.6.2, we outline its practical implementation.

6.6.1 Construction of the augmented permutation test

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ denote the $(n \times 1)$-dimensional vector of outcomes, $\boldsymbol{A} = (A_1, \ldots, A_n)^T$ the $(n \times 1)$-dimensional vector of treatment assignments and $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ the $(n \times p)$-matrix of collected covariates on the n individuals. The aim is to derive the permutation null distribution of the locally efficient estimator (6.17), treating the number of treated N_1 and the outcome and covariate data $(\boldsymbol{Y}, \mathbb{X})$ as fixed. Let $\mathcal{G}$ denote the set of all permutations of $\{1, \ldots, n\}$ that considers all $M_n = \binom{n}{N_1}$ partitions of $\{1, \ldots, n\}$ into two groups (one of size N_1 and the other of size N_0). For an arbitrary $g \in \mathcal{G}$ and for an arbitrary $(n \times 1)$-dimensional vector $\boldsymbol{W} = (W_1, \ldots, W_n)^T$ or for an arbitrary $(n \times p)$-matrix $\mathbb{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)^T$ with rows $\boldsymbol{W}_i^T$, let $g(\boldsymbol{W}) = (W_{g(1)}, \ldots, W_{g(n)})^T$, $g(\mathbb{W}) = (\boldsymbol{W}_{g(1)}, \ldots, \boldsymbol{W}_{g(n)})^T$ respectively, with $g(i)$ the ith element of the permutation g. The null hypothesis $A \perp\!\!\!\perp (Y, \boldsymbol{X})$ implies that

$$(g(\boldsymbol{Y}), \boldsymbol{A}, g(\mathbb{X})) \stackrel{d}{=} (\boldsymbol{Y}, \boldsymbol{A}, \mathbb{X}) \stackrel{d}{=} (\boldsymbol{Y}, g(\boldsymbol{A}), \mathbb{X}) \tag{6.20}$$

for all $g \in \mathcal{G}$ (also called the **randomization hypothesis** (Lehmann and Romano 2005)). The permutation null distribution of (6.17) can then be obtained by re-calculating (6.17) for every permutation $g(\boldsymbol{A})$ since the null distribution of (6.17) is invariant under such permutations. This involves refitting the working model $m(0, 1, \boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau})$ for every $g \in \mathcal{G}$ since the estimator $\hat{\boldsymbol{\tau}}_{n,g}$ of $\boldsymbol{\tau}$ under permutation g depends on $g(\boldsymbol{A})$. When this is computationally cumbersome, then greater computational efficiency can be obtained by noting that the null hypothesis implies that the CPI $P(Y_i \preceq Y_j | A_i, A_j, \boldsymbol{X}_i, \boldsymbol{X}_j)$ equals $P(Y_i \preceq Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j)$, which is no longer a function of $g(\boldsymbol{A})$, and thus remains fixed throughout the permutation procedure. The latter can be estimated using a working model $m_0(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau})$. We will therefore work with the test statistic

$$\begin{aligned} T(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X}; \boldsymbol{\tau}) = \frac{1}{N_0 N_1} \sum_{i=1}^{n} \sum_{j \neq i} (1 - A_i) A_j I(Y_i \preceq Y_j) \\ + \sum_{i=1}^{n} \sum_{j \neq i} \left\{ \frac{1}{n(n-1)} - \frac{(1 - A_i) A_j}{N_1 N_0} \right\} m_0(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau}). \end{aligned}$$

An estimator $\hat{\boldsymbol{\tau}}_n(\mathbb{X}, \boldsymbol{Y})$ of $\boldsymbol{\tau}$, obtained by solving the estimating equations (6.22), now only depends on the outcomes $\boldsymbol{Y}$ and the covariates $\mathbb{X}$, but not on the treatment assignments $\boldsymbol{A}$. As such, we obtain the test statistic

$$\widehat{T}_n(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X}) \equiv T(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X}; \hat{\boldsymbol{\tau}}_n(\mathbb{X}, \boldsymbol{Y})).$$

The working model m_0 used in the construction of $\widehat{T}_n$ is different from the working model m used in the construction of $\hat{v}_{n,\text{adap}}$. When $A \perp\!\!\!\perp Y|\boldsymbol{X}$ is false, m_0 is a misspecified working model for the CPI $P(Y_i \preceq Y_j|A_i, A_j, \boldsymbol{X}_i, \boldsymbol{X}_j)$. This implies that $\hat{v}_{n,\text{adap}}$ may be asymptotically more efficient than $\widehat{T}_n$ under model $\mathcal{M}_{\text{indep}}$ and hence, one may expect the permutation test based on $\widehat{T}_n$ to have lower power than the permutation test based on $\hat{v}_{n,\text{adap}}$. Simulation studies in Section 6.8 will compare the performance of both permutation tests based on $\widehat{T}_n$ and $\hat{v}_{n,\text{adap}}$ respectively.

The permutation null distribution of the test statistic $\widehat{T}_n(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X})$ (with the number of treated subjects N_1 and the observed data $(\boldsymbol{Y}, \mathbb{X})$ fixed) is now obtained by calculating $\widehat{T}_n(g(\boldsymbol{A}), \boldsymbol{Y}, \mathbb{X})$ for all M_n permutations $g \in \mathcal{G}$. It is a discrete distribution taking values $\widehat{T}_n(g(\boldsymbol{A}), \boldsymbol{Y}, \mathbb{X})$, $g \in \mathcal{G}$, with probability the frequency that value occurred among the permutations. From this, we can obtain a conditional level α test. Furthermore, interestingly, we show that the outlined procedure also has unconditional level α. The development below is a modification of the argument due to Hoeffding (1952). Let the sequence

$$\widehat{T}_n^{(1)} \leq \widehat{T}_n^{(2)} \leq \ldots \leq \widehat{T}_n^{(M_n-1)} \leq \widehat{T}_n^{(M_n)}$$

denote the order statistics of the set $\{\widehat{T}_n(g(\boldsymbol{A}), \boldsymbol{Y}, \mathbb{X}) : g \in \mathcal{G}\}$. Given a nominal significance level α, $0 < \alpha < 1$, define the numbers $\ell = \lfloor M_n\alpha/2 \rfloor$ and $k = M_n - \lfloor M_n\alpha/2 \rfloor$. Let $M^+(\mathbb{X}, \boldsymbol{Y}) = \sum_{i=1}^{M_n} I(\widehat{T}_n^{(i)} > \widehat{T}_n^{(k)})$, $M_0^+(\mathbb{X}, \boldsymbol{Y}) = \sum_{i=1}^{M_n} I(\widehat{T}_n^{(i)} = \widehat{T}_n^{(k)})$, $M^-(\mathbb{X}, \boldsymbol{Y}) = \sum_{i=1}^{M_n} I(\widehat{T}_n^{(i)} < \widehat{T}_n^{(\ell)})$, $M_0^-(\mathbb{X}, \boldsymbol{Y}) = \sum_{i=1}^{M_n} I(\widehat{T}_n^{(i)} = \widehat{T}_n^{(\ell)})$ and set

$$a^+(\mathbb{X}, \boldsymbol{Y}) = \frac{M_n\alpha/2 - M^+(\mathbb{X}, \boldsymbol{Y})}{M_0^+(\mathbb{X}, \boldsymbol{Y})} \qquad \text{and} \qquad a^-(\mathbb{X}, \boldsymbol{Y}) = \frac{M_n\alpha/2 - M^-(\mathbb{X}, \boldsymbol{Y})}{M_0^-(\mathbb{X}, \boldsymbol{Y})}.$$

Because $M^+(\mathbb{X}, \boldsymbol{Y}) \leq M_n - k \leq M_n\alpha/2$ and $M^+(\mathbb{X}, \boldsymbol{Y}) + M_0^+(\mathbb{X}, \boldsymbol{Y}) \geq M_n - k + 1 > M_n\alpha/2$, we have $0 \leq a^+(\mathbb{X}, \boldsymbol{Y}) < 1$ and similarly $0 \leq a^-(\mathbb{X}, \boldsymbol{Y}) < 1$. An exact

permutation test can formally be defined using the test function

$$\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = \begin{cases} 1 & \text{if } \widehat{T}_n(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) > \widehat{T}_n^{(k)} \text{ or } \widehat{T}_n(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) < \widehat{T}_n^{(\ell)}, \\ a^+(\mathbb{X},\boldsymbol{Y}) & \text{if } \widehat{T}_n(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = \widehat{T}_n^{(k)}, \\ a^-(\mathbb{X},\boldsymbol{Y}) & \text{if } \widehat{T}_n(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = \widehat{T}_n^{(\ell)}, \\ 0 & \text{if } \widehat{T}_n^{\ell} < \widehat{T}_n(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) < \widehat{T}_n^{(k)}. \end{cases}$$

The test based on the test function Ψ rejects H_0 if $\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = 1$ in favor of the alternative $P(Y \preceq Y^*|A = 0, A^* = 1) \neq 0.5$ and the test does not reject H_0 if $\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = 0$. When $\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = a^+(\mathbb{X},\boldsymbol{Y})$ ($\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = a^-(\mathbb{X},\boldsymbol{Y})$), H_0 is randomly rejected with probability $a^+(\mathbb{X},\boldsymbol{Y})$ ($a^-(\mathbb{X},\boldsymbol{Y})$). By definition, the test defined by Ψ has exact conditional level α. In practice, the test may be defined by only rejecting when $\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y}) = 1$ leading to a test only with conditional level that may be slightly lower than α due to the discreteness of the permutation null distribution.

The following theorem shows that the test based on Ψ also has unconditional level α, which follows from the invariance property (6.20), implied by $H_0 : A \perp\!\!\!\perp (Y, \boldsymbol{X})$. The original result is due to Hoeffding (1952).

Theorem 6.5. *For the observed data $(\boldsymbol{Y},\boldsymbol{A},\mathbb{X})$, obtained by i.i.d. sampling from a larger super-population, and given the invariance property (6.20) implied by $H_0 : A \perp\!\!\!\perp (Y,\boldsymbol{X})$, the permutation test defined by the test function Ψ has unconditional level α.*

Proof. By definition of a^+, a^- and Ψ, we have

$$\begin{aligned} &\frac{1}{M_n}\sum_{g\in\mathcal{G}} \Psi(g(\boldsymbol{A}),\mathbb{X},\boldsymbol{Y}) \\ &\quad = \frac{M^+(\mathbb{X},\boldsymbol{Y}) + M^-(\mathbb{X},\boldsymbol{Y}) + a^+(\mathbb{X},\boldsymbol{Y})M_0^+(\mathbb{X},\boldsymbol{Y}) + a^-(\mathbb{X},\boldsymbol{Y})M_0^-(\mathbb{X},\boldsymbol{Y})}{M_n} \\ &\quad = \alpha. \end{aligned}$$

Below, we use the subscript H_0 to indicate that probabilities and expectations are

calculated under H_0. The invariance property (6.20) implies that

$$E_{H_0}\{\Psi(g(\boldsymbol{A}),\mathbb{X},\boldsymbol{Y})|N_1\} = E_{H_0}\{\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y})|N_1\}$$

for all $g \in \mathcal{G}$ and thus

$$\begin{aligned} P_{H_0}(\text{reject } H_0|N_1) &= E_{H_0}\{\Psi(\boldsymbol{A},\mathbb{X},\boldsymbol{Y})|N_1\} = \frac{1}{M_n}\sum_{g\in\mathcal{G}} E_{H_0}\{\Psi(g(\boldsymbol{A}),\mathbb{X},\boldsymbol{Y})|N_1\} \\ &= E_{H_0}\left\{\frac{1}{M_n}\sum_{g\in\mathcal{G}}\Psi(g(\boldsymbol{A}),\mathbb{X},\boldsymbol{Y})\Big|N_1\right\} = \alpha \end{aligned}$$

and consequently also unconditionally. □

6.6.2 Implementation of the augmented permutation test

The proposed procedure can be implemented using the following steps:

Step 1. Fit a working PIM $m_0(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})$ for the CPI $P(Y_i \preceq Y_j|\boldsymbol{X}_i,\boldsymbol{X}_j)$ based on the outcome and covariate data $(\boldsymbol{Y},\mathbb{X})$.

Step 2. Using the working PIM from Step 1, calculate the observed test statistic $\widehat{T}_{n,0} \equiv \widehat{T}_n(\boldsymbol{A},\boldsymbol{Y},\mathbb{X})$ based on the original data $(\boldsymbol{Y},\boldsymbol{A},\mathbb{X})$.

Step 3. Given the observed data $(\boldsymbol{Y},\mathbb{X})$ and the number of treated subjects N_1, reassign subjects to treatment using the permuted vector of treatment assignments $g(\boldsymbol{A})$ and calculate $\widehat{T}_{n,g} \equiv \widehat{T}_n(g(\boldsymbol{A}),\boldsymbol{Y},\mathbb{X})$ for all M_n permutations $g \in \mathcal{G}$, with $\boldsymbol{A}$ the original treatment assignment.

Step 4. Order all M_n values $\widehat{T}_{n,g}$ of the permutation null distribution and obtain the sequence of ordered statistics

$$\widehat{T}_n^{(1)} \leq \widehat{T}_n^{(2)} \leq \ldots \leq \widehat{T}_n^{(M_n-1)} \leq \widehat{T}_n^{(M_n)}.$$

Step 5. Reject H_0 if $\widehat{T}_{n,0} < \widehat{T}_n^{(\ell)}$ or $\widehat{T}_{n,0} > \widehat{T}_n^{(k)}$ with ℓ the largest integer such that $P_{\text{per},H_0}(\widehat{T}_n < \widehat{T}_n^{(\ell)}) \leq \alpha/2$ and with k the smallest integer such that $P_{\text{per},H_0}(\widehat{T}_n > \widehat{T}_n^{(k)}) \leq \alpha/2$ where the subscript per, H_0 is used to indicate

that the probabilities are defined with respect to the permutation null distribution. The exact conditional (two-sided) p-value can be obtained as

$$\tilde{p} = 2 \times \min \left\{ \frac{1}{M_n} \sum_{i=1}^{M_n} I(\widehat{T}_n^{(i)} \leq \widehat{T}_{n,0}), \frac{1}{M_n} \sum_{i=1}^{M_n} I(\widehat{T}_n^{(i)} \geq \widehat{T}_{n,0}) \right\}.$$

The null hypothesis can then be equivalently rejected if $\tilde{p} \leq \alpha$.

Conditional on $(\boldsymbol{Y}, \mathbb{X})$ and the number of treated subjects N_1, this testing procedure has conditional level α by definition of k and ℓ. However, by Theorem 6.5, it follows that when $\{\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i) : i = 1, \ldots, n\}$ is obtained by i.i.d. sampling from a larger super-population, this testing procedure also has unconditional level $P_{H_0}(\text{reject } H_0) = \alpha$, where the subscript H_0 is used to indicate that the probabilities are calculated under H_0. Furthermore, the resulting testing procedure is consistent for the alternative that the MPI $P(Y \preceq Y^* | A = 0, A^* = 1) \neq 0.5$ because $\widehat{T}_n(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X})$ is based on a consistent estimator for the true MPI (van der Vaart 1998, Lemma 14.15). The latter holds whether or not we include treatment in the working model for the CPI.

In practice, the total number of permutations M_n may become very large, e.g., for $n = 30$ and $N_1 = 15$, $M_n \approx 1.5 \times 10^8$. Hence, it may be infeasible to list all possible permutations. Alternatively, one could randomly sample B of the possible permutations with replacement to approximate the permutation null distribution with arbitrary accuracy by increasing the number of samples B. The testing procedure outlined above then remains exactly the same with the only difference that M_n is replaced by B. Instead of an exact p-value $\tilde{p}$, an approximate p-value $\hat{p}_B$ is obtained, whose accuracy can be measured by means of a 95% CI $[\hat{p}_B \pm 1.96 \hat{p}_B(1 - \hat{p}_B)/\sqrt{B}]$. When the accuracy is not sufficient, the number of random permutations B can be increased. In Appendix 6.D.2, we provide an R-function that implements the permutation test.

We end this section by noting that, interestingly, variable selection procedures in a model for the CPI $P(Y_i \preceq Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j)$ can be safely used in the construction of the test statistic $\widehat{T}_n(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X})$ without compromising the Type I error rate. This is because under the null hypothesis (6.19), treatment does not need to be included in

the working model for the CPI, thereby making $m(\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\boldsymbol{\tau}}_n(\mathbb{X}, \boldsymbol{Y}))$ a function of $(\boldsymbol{Y}, \mathbb{X})$ only. Consequently, the variable selection procedure remains fixed for all permutations, assuring conditional level α. From the proof of Theorem 6.5, it then also follows that the permutation test continues to have unconditional level α upon averaging over the joint distribution of the observed data and hence also over the variable selection procedure which is a function of $(\boldsymbol{Y}, \mathbb{X})$ only, as in Stephens et al. (2013).

6.7 Data Analysis: ACTG 175

We reanalyze data from the AIDS Clinical Trials Group Protocol 175 (ACTG 175), which randomized $n = 2139$ HIV-infected patients (double-blind) to four different antiretroviral regimens in equal proportions: zidovudine (ZDV) monotherapy, ZDV plus didanosine (ddI), ZDV plus zalcitabine, and ddI monotherapy (Hammer et al. 1996). We follow the analyses in Leon et al. (2003); Davidian et al. (2005); Tsiatis et al. (2008) and consider two treatment groups (A): ZDV monotherapy ($A = 0$, $N_0 = 532$) versus the other three treatment regimens combined ($A = 1$, $N_1 = 1607$), resulting in a randomization probability $\pi = 0.75$. We are interested in the effect of treatment A on CD4 count (cells/mm^3) at 20 ± 5 weeks post-baseline (Y). Extensive baseline information was collected. We consider the continuous measurements baseline CD4 count (cells/mm^3, X_1), baseline CD8 count (cells/mm^3, X_2), age (years, X_3), weight (kg, X_4), Karnofsky score (on a $0-100$ scale, X_5), and the binary indicator variables for hemophilia (X_6), homosexual activity (X_7), history of intravenous drug use (X_8), race (X_9), gender (X_{10}), antiretroviral history (X_{11}), and symptomatic status (X_{12}).

Figure 6.1 shows histograms of CD4 count at 20 ± 5 weeks post-baseline for both treatment groups and Figure 6.2 shows QQ-plots of CD4 count at 20 ± 5 weeks post-baseline for both treatment groups. Both outcome distributions are modestly skewed to the right. Hence, expressing the treatment effect on the scale of the probabilistic index may be indicated. An estimate for the MPI (calculated as U/N_0N_1) equals 0.586 ($p < 0.001$) indicating strong evidence that the true MPI differs from 0.5.

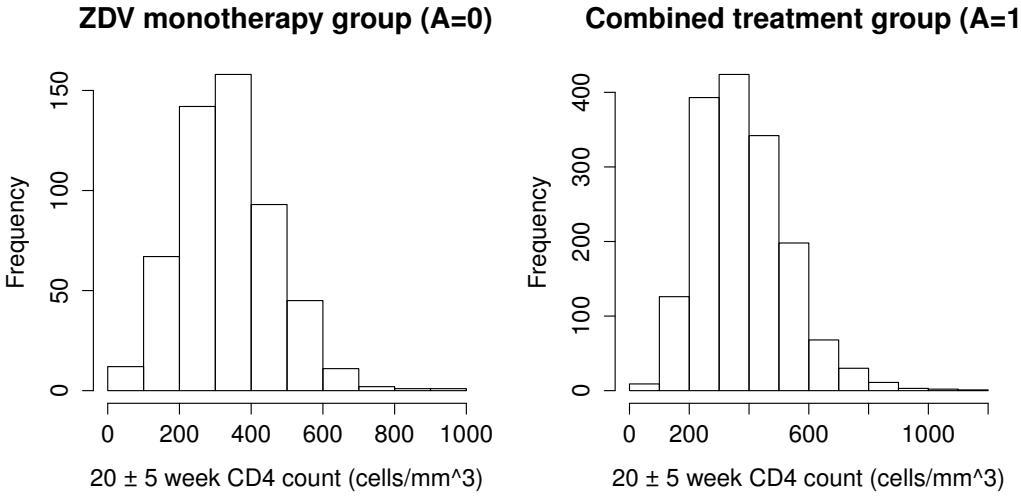


Figure 6.1: *Histograms of CD4 counts at* 20 ± 5 *weeks post-baseline: ZDV monotherapy group (*$A = 0$*, left) and combined therapy group (*$A = 1$*, right).*

To gain insight into the performance of the proposed approach, we analyze 1000 random subsamples of size $n_{\text{sub}} = 30, 50, 100$ of the original dataset using the various methods. In particular, we apply the classical Mann-Whitney U test and the proposed tests based on $B = 10000$ random permutations and working PIMs using both a probit and logit link adjusting for (i) baseline CD4 count only (labeled BASE), (ii) those covariates that have a significant marginal association with the outcome at the 5% significance level in a PIM with the same link function as the corresponding working PIM (labeled SIG).

The power of the different testing procedures is shown in Table 6.1. Table 6.2 shows the percentage of times that each variable is included in the working PIMs in the variable selection procedure. From the relative efficiencies, we observe a substantial decrease in variance of the augmented Mann-Whitney U statistics versus the ordinary Mann-Whitney U statistic, roughly between 40% and 50%, in spite of the working PIMs being possibly misspecified. This means that to obtain the same power, we need between 40% and 50% individuals less in the study. Table 6.1 confirms that this increase in efficiency by incorporating the baseline covariate information indeed translates into a major power gain. This decrease in variance is also illustrated in Figure 6.3, where we observe that the variance of the

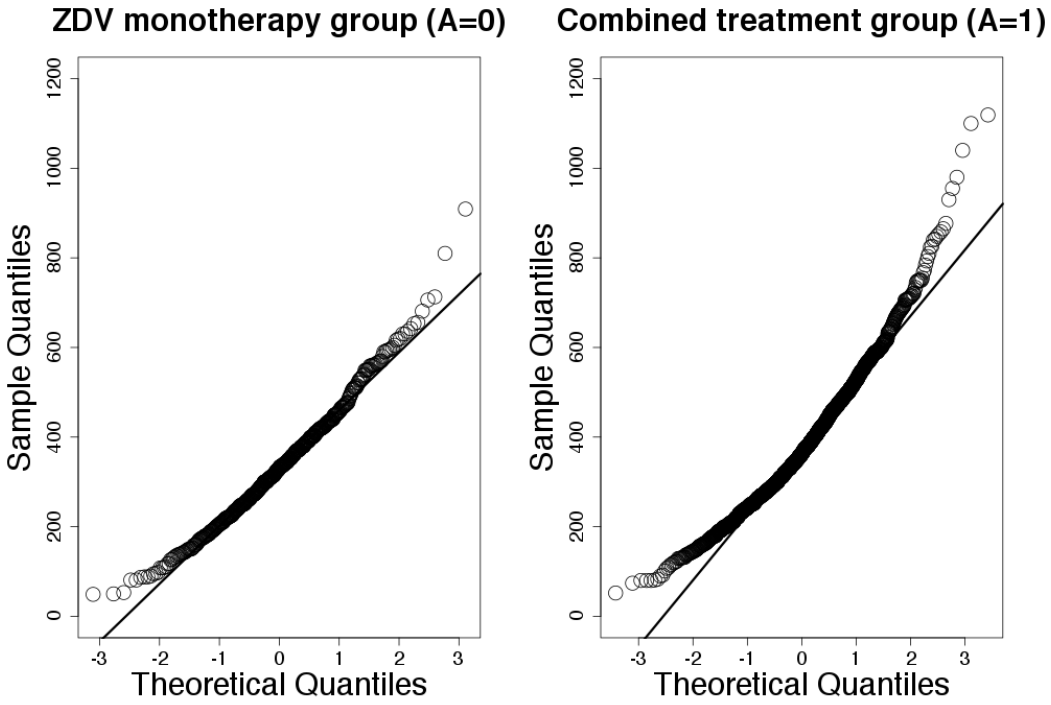


Figure 6.2: *QQ-plots of CD4 counts at* 20 ± 5 *weeks post-baseline: ZDV monotherapy group (*$A = 0$*, left) and combined therapy group (*$A = 1$*, right).*

permutation null distribution of the augmented Mann-Whitney U statistic (p-value of 0.017, based on a working PIM fitted under the null hypothesis using logit link, with adjustment for baseline CD4) has lower variance than that of the standard Mann-Whitney U test statistic (p-value of 0.282), both obtained from a random subsample of size $n_{\text{sub}} = 50$ of the original dataset.

6.8 Simulation Studies

We report several simulation studies to demonstrate the performance of the proposed estimator and Wald test (outlined in Section 6.5) and the proposed permutation test (outlined in Section 6.6) as compared to the classical Mann-Whitney U test, each involving 1000 Monte Carlo runs.

Table 6.1: *Data analysis on* 1000 *random subsamples of the original ACTG 175 data set.*

TEST	POWER	RE	POWER	RE	POWER	RE
	$n_{\text{sub}} = 30$		$n_{\text{sub}} = 50$		$n_{\text{sub}} = 100$	
MW	9.7	1.00	15.3	1.00	25.8	1.00
augMW probit0 BASE	14.9	0.60	23.1	0.56	44.9	0.59
augMW probit0 SIG	14.0	0.54	22.9	0.51	44.3	0.56
augMW logit0 BASE	15.0	0.60	23.2	0.56	45.1	0.59
augMW logit0 SIG	14.1	0.55	23.6	0.51	44.8	0.56

NOTE: RE: empirical variance of the augmented test statistic $\widehat{T}_n(\boldsymbol{A}, \boldsymbol{Y}, \mathbb{X})$ divided by the empirical variance of $U/(N_1 N_0)$; BASE: adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model; MW: Mann-Whitney U test; augMW: augmented Mann-Whitney U test; probit0: working PIM fitted under the null hypothesis using probit link; logit0: working PIM fitted under the null hypothesis using logit link.

6.8.1 Data generation

The data generating mechanism is based on the fit of the ACTG 175 data from Tsiatis et al. (2008), Section 5, also used in Section 6.7. For each simulated data set, we generated the continuous baseline covariates CD4 count (X_1), CD8 count (X_2), age (X_3), weight (X_4) and Karnofsky score (X_5) from a multivariate normal distribution with mean

$$(350.5, 986.6, 35.2, 75.1, 95.4)$$

and covariance matrix

$$\begin{bmatrix} 14059.8 & 12200.5 & -41.6 & 57.2 & 54.4 \\ & 230589.9 & 196.0 & 573.7 & -24.3 \\ & & 75.8 & 15.3 & -5.1 \\ & & & 175.9 & 2.7 \\ & & & & 34.8 \end{bmatrix},$$

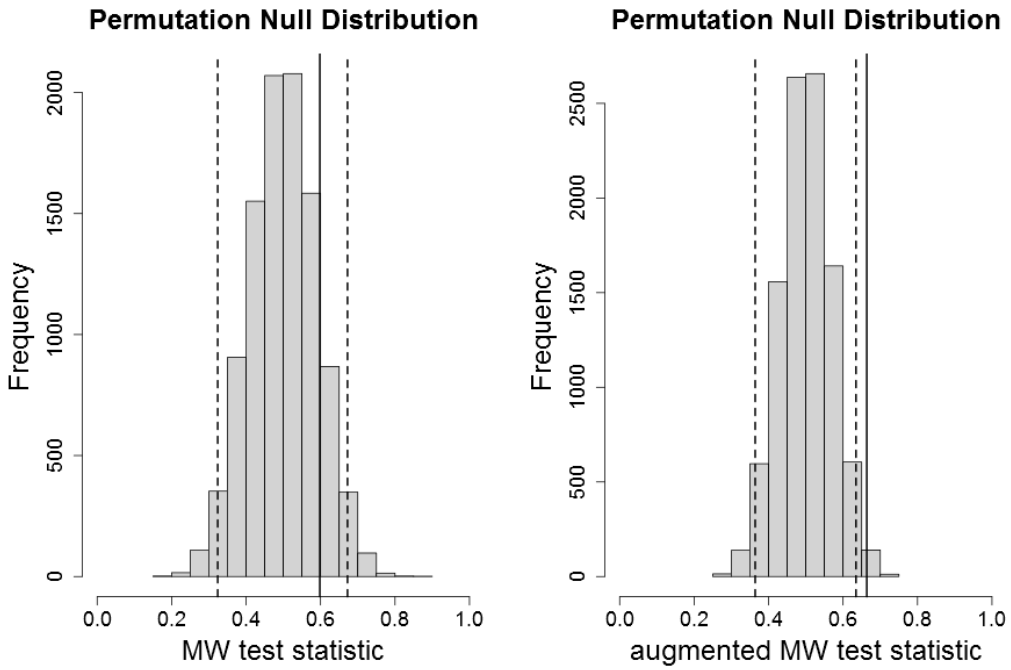


Figure 6.3: *Permutation null distribution of the standard Mann-Whitney U statistic (left) and the augmented Mann-Whitney U statistic (based on a working PIM fitted under the null hypothesis using logit link, with adjustment for baseline CD4, right) each based on a random subsample of size $n_{sub} = 50$, showing the rejection region (dashed) and the observed test statistic (solid).*

the empirical mean and covariance matrix of these variables in the ACTG 175 data. Next, independently, baseline binary indicators were generated for hemophilia (X_6), homosexual activity (X_7), history of drug use (X_8), race (X_9), gender (X_{10}), antiretroviral history (X_{11}) and symptomatic status (X_{12}) from independent Bernoulli distributions with proportions

$$(0.08, 0.66, 0.13, 0.29, 0.83, 0.59, 0.17),$$

equalling the empirical proportions of these variables in the ACTG 175 data. A treatment indicator A was generated from a Bernoulli distribution with probability of being treated $\pi = 0.5$ and 0.75, indicating a balanced and unbalanced design, independently from the covariates $\boldsymbol{X} = (X_1, \ldots, X_{12})$. Finally, the outcome CD4 count at 20 ± 5 weeks (Y) was generated as $\mu + \sigma T_3$, with T_3 a student t-distribution

Table 6.2: *Percentage variables included in the PIMs for the SIG strategy for the* 1000 *random subsamples of the original ACTG 175 data set.*

COV	$n_{sub} = 30$	$n_{sub} = 50$	$n_{sub} = 100$	$n_{sub} = 30$	$n_{sub} = 50$	$n_{sub} = 100$
	probit PIM			logit PIM		
X_1	91.7	98.1	100.0	90.5	97.7	100.0
X_2	14.9	12.5	10.1	13.3	10.8	9.3
X_3	11.3	8.9	7.3	10.5	8.6	6.5
X_4	12.7	7.5	8.0	11.7	7.1	7.3
X_5	15.1	14.5	22.6	14.0	13.5	22.1
X_6	23.3	21.2	16.9	22.3	20.3	15.7
X_7	8.4	7.1	6.5	7.7	6.7	6.4
X_8	19.0	12.5	6.9	18.2	11.6	6.0
X_9	10.5	9.1	5.6	10.0	8.6	5.3
X_{10}	14.6	9.6	6.2	13.4	8.7	5.9
X_{11}	31.6	40.0	66.9	30.1	39.2	66.3
X_{12}	22.3	19.0	24.9	21.0	18.0	24.0

NOTE: COV, covariate; logit, working PIM fitted using logit link; probit, working PIM fitted using probit link.

with three degrees of freedom, $\mu(\boldsymbol{X}) = (1-\pi)\mu_0(\boldsymbol{X}) + \pi\mu_1(\boldsymbol{X})$ under H_0 and $\mu(A,\boldsymbol{X}) = (1-A)\mu_0(\boldsymbol{X}) + A\mu_1(\boldsymbol{X})$ under the alternative where

$$\begin{aligned}
\mu_0(\boldsymbol{X}) =& -79.705 + 1.599X_1 - 0.0007X_1^2 \\
& - 0.107X_1X_6 - 0.005X_1X_4 + 0.013X_4X_5 - 0.040X_2X_{11} - 23.199X_7X_9, \\
\mu_1(\boldsymbol{X}) =& 95.445 + 1.1X_1 - 0.0005X_1^2 - 142.288X_7 - 0.178X_1X_8 - 0.087X_1X_9 \\
& + 0.033X_2X_6 - 0.014X_2X_7 - 0.021X_2X_{11} - 0.72X_3X_{11} - 0.554X_3X_{12} \\
& - 0.706X_4X_6 + 1.282X_4X_8 + 1.688X_5X_7 - 28.321X_8X_9 \\
& - 45.337X_8X_{10} + 35.981X_8X_{11} + 24.032X_9X_{11} - 3.602X_{10}X_{11},
\end{aligned}$$

$\sigma = (1-\pi)95.82 + \pi 115.63$ under H_0 and $\sigma(A) = (1-A)95.82 + A115.63$ under the alternative. A t-distribution with three degrees of freedom is used to induce severe outliers in the outcome distribution, which lends itself to expressing treatment effects on the probabilistic index (PI) scale rather than the linear scale. With $\beta_0 = E(Y|A=1) - E(Y|A=0)$ and $\nu_0 = P(Y \preceq Y^*|A=0, A^*=1)$, the true value

is given by $\beta_0 = 0$ and $\nu_0 = 0.5$ under H_0 and $\beta_0 = 63$ and $\nu_0 = 0.606$ under the alternative.

6.8.2 Estimation

For each data set, the MPI ν_0 was first estimated using the unadjusted estimator U/N_0N_1 (MW). Next, ν_0 was estimated using the locally efficient estimator $\hat{\nu}_{n,\text{adap}}$ as presented in (6.17) (augMW) using several working models $m(A_i, A_j, \boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau})$ for the CPI: both a logit and probit link function were used with on the one hand adjustment only for baseline CD4 count (X_1), labeled BASE throughout, and on the other hand adjustment for those covariates showing a significant marginal association with the outcome in the corresponding working model, labeled SIG throughout. Table 6.10 and Table 6.11 in Appendix 6.E show the percentage of times of all 1000 Monte Carlo replications that each variable was included in the working PIMs for both the probit and logit link function (for sample sizes $n = 50$ and 100). Results for the case $\nu_0 = 0.606$ for sample sizes $n = 50$, 100 and 200 are shown in Table 6.3 ($\pi = 0.5$) and Table 6.4 ($\pi = 0.75$); results (albeit similar) for the case $\nu_0 = 0.5$ for the same sample sizes are shown in Table 6.5 and Table 6.6.

All estimators are unbiased. Furthermore, the sandwich estimator for the standard error calculated via (6.18) succeeds quite well in capturing the finite-sample variability of the estimators, resulting in good coverage of the 95% Wald confidence intervals. Some minor undercoverage is observed for the unbalanced case $\pi = 0.75$. Substantial gains in efficiency are observed for all augmented estimators by exploiting the covariate information as compared to the unadjusted estimator. For this particular setting, reductions in variance range between 15% and 25%. Note that this does not come at the cost of bias due to model misspecification.

6.8.3 Testing

To test for the absence of a treatment effect, we consider a two-sample t-test and we also test for $\beta_A = 0$ in the linear regression (LR) model (6.3) with adjustment strategies BASE and SIG. Next, we consider a Wald test of $\tau_A = 0$ as in the PIM (6.5) but using both a logit and probit link (BASE and SIG). Next, we also show

Table 6.3: *Simulation results for estimation of the MPI* $\nu_0 = P(Y \preceq Y^* | A = 0, A^* = 1) = 0.606$, $\pi = 0.5$, *based on* 1000 *Monte Carlo replications.*

ESTIMATOR	BIAS	RMSE	MCSD	AVESE	COV	RE
			$\pi = 0.5$			
			$n = 50$			
MW	−0.0044	0.080	0.080	0.082	0.94	1.00
augMW probit BASE	−0.0005	0.069	0.069	0.072	0.95	0.74
augMW probit SIG	−0.0061	0.068	0.068	0.069	0.95	0.72
augMW logit BASE	−0.0006	0.069	0.069	0.071	0.95	0.74
augMW logit SIG	−0.0060	0.068	0.068	0.069	0.95	0.72
			$n = 100$			
MW	0.0020	0.055	0.055	0.057	0.95	1.00
augMW probit BASE	0.0016	0.049	0.049	0.050	0.95	0.79
augMW probit SIG	−0.0011	0.049	0.049	0.049	0.95	0.79
augMW logit BASE	0.0015	0.049	0.049	0.050	0.95	0.79
augMW logit SIG	−0.0010	0.049	0.049	0.049	0.95	0.79
			$n = 200$			
MW	−0.0006	0.040	0.040	0.040	0.95	1.00
augMW probit BASE	−0.0006	0.036	0.036	0.035	0.94	0.81
augMW probit SIG	−0.0020	0.036	0.036	0.036	0.94	0.81
augMW logit BASE	−0.0007	0.036	0.036	0.036	0.94	0.81
augMW logit SIG	−0.0019	0.036	0.036	0.036	0.94	0.81

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; AVESE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; RE, relative efficiency is the Monte Carlo variance of the indicated estimator divided by the Monte Carlo variance of $U/(N_0 N_1)$; MW, unadjusted estimator based on the Mann-Whitney U test statistic; augMW: augmented MW; probit, working PIM fitted using probit link; logit, working PIM fitted using logit link; BASE, adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model.

results for the classical Mann-Whitney U test based on asymptotic approximations and for the Wald tests based on the asymptotic distribution of the test statistic $Z = (\hat{\nu}_{n,\text{adap}} - 0.5)/\{\widehat{\text{var}}_n(\hat{\nu}_{n,\text{adap}})\}^{1/2}$ for the estimators considered in Section 6.5. Finally, we also show results for the tests presented in Zhang et al. (2008), Section 5, labeled ZHANG (both unadjusted and adjusted), and for the permutation methods

Table 6.4: *Simulation results for estimation of the MPI* $\nu_0 = P(Y \preceq Y^*|A = 0, A^* = 1) = 0.606$, $\pi = 0.75$, *based on* 1000 *Monte Carlo replications.*

ESTIMATOR	BIAS	RMSE	MCSD	AVESE	COV	RE
			$\pi = 0.75$			
			$n = 50$			
MW	0.0004	0.089	0.089	0.091	0.95	1.00
augMW probit BASE	−0.0019	0.081	0.081	0.079	0.94	0.83
augMW probit SIG	−0.0093	0.081	0.080	0.077	0.94	0.81
augMW logit BASE	−0.0019	0.081	0.081	0.079	0.94	0.83
augMW logit SIG	−0.0088	0.081	0.081	0.076	0.94	0.83
			$n = 100$			
MW	−0.0001	0.064	0.064	0.064	0.94	1.00
augMW probit BASE	0.0001	0.055	0.055	0.055	0.94	0.74
augMW probit SIG	−0.0026	0.054	0.054	0.054	0.94	0.71
augMW logit BASE	0.0006	0.055	0.055	0.055	0.94	0.74
augMW logit SIG	−0.0025	0.055	0.055	0.054	0.94	0.74
			$n = 200$			
MW	−0.0017	0.046	0.046	0.045	0.93	1.00
augMW probit BASE	−0.0010	0.040	0.040	0.039	0.93	0.76
augMW probit SIG	−0.0022	0.040	0.040	0.038	0.94	0.76
augMW logit BASE	−0.0010	0.040	0.040	0.039	0.93	0.76
augMW logit SIG	−0.0022	0.040	0.040	0.038	0.94	0.76

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; AVESE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; RE, relative efficiency is the Monte Carlo variance of the indicated estimator divided by the Monte Carlo variance of $U/(N_0N_1)$; MW, unadjusted estimator based on the Mann-Whitney U test statistic; augMW: augmented MW; probit, working PIM fitted using probit link; logit, working PIM fitted using logit link; BASE, adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model.

outlined in Section 6.6. We consider both the classical Mann-Whitney U test and the proposed method where we augment the Mann-Whitney U test statistic with both a working logit and probit PIM (BASE and SIG). For illustrative purposes only, we do not only present results excluding treatment in the working model (labeled 0) for the CPI, but we also present results using a working model including

Table 6.5: *Simulation results for estimation of the MPI* $\nu_0 = P(Y \preceq Y^*|A = 0, A^* = 1) = 0.5$, $\pi = 0.5$, *based on* 1000 *Monte Carlo replications.*

ESTIMATOR	BIAS	RMSE	MCSD	AVESE	COV	RE
			$\pi = 0.5$			
			$n = 50$			
MW	−0.0053	0.081	0.081	0.084	0.95	1.00
augMW probit BASE	−0.0013	0.070	0.070	0.073	0.95	0.76
augMW probit SIG	−0.0005	0.068	0.068	0.070	0.95	0.69
augMW logit BASE	−0.0013	0.070	0.070	0.073	0.95	0.76
augMW logit SIG	−0.0004	0.068	0.068	0.071	0.95	0.71
			$n = 100$			
MW	0.0016	0.057	0.057	0.059	0.95	1.00
augMW probit BASE	0.0012	0.050	0.050	0.051	0.96	0.75
augMW probit SIG	0.0016	0.050	0.050	0.050	0.95	0.72
augMW logit BASE	0.0012	0.050	0.050	0.051	0.96	0.75
augMW logit SIG	0.0015	0.050	0.050	0.050	0.95	0.72
			$n = 200$			
MW	−0.0010	0.040	0.040	0.041	0.95	1.00
augMW probit BASE	−0.0010	0.036	0.036	0.036	0.94	0.77
augMW probit SIG	−0.0008	0.036	0.036	0.035	0.94	0.73
augMW logit BASE	−0.0010	0.036	0.036	0.036	0.94	0.77
augMW logit SIG	−0.0008	0.036	0.036	0.035	0.94	0.73

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; AVESE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; RE, relative efficiency is the Monte Carlo variance of the indicated estimator divided by the Monte Carlo variance of $U/(N_0N_1)$; MW, unadjusted estimator based on the Mann-Whitney U test statistic; augMW: augmented MW; probit, working PIM fitted using probit link; logit, working PIM fitted using logit link; BASE, adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model.

treatment (labeled 1). All permutation null distributions (for both the classical Mann-Whitney U test and the augmented Mann-Whitney U tests) are approximated using $B = 10000$ random permutations. Results for the (unconditional) Type I error and power (and relative efficiency) for the considered tests for $\pi = 0.5$ and 0.75 are shown in Table 6.7 ($n = 50$) and Table 6.8 ($n = 100$). For those methods involving

Table 6.6: *Simulation results for estimation of the MPI* $\nu_0 = P(Y \preceq Y^*|A = 0, A^* = 1) = 0.5$, $\pi = 0.75$, *based on* 1000 *Monte Carlo replications.*

ESTIMATOR	BIAS	RMSE	MCSD	AVESE	COV	RE
			$\pi = 0.75$			
			$n = 50$			
MW	−0.0004	0.095	0.095	0.096	0.94	1.00
augMW probit BASE	−0.0028	0.086	0.086	0.085	0.94	0.78
augMW probit SIG	−0.0042	0.085	0.086	0.082	0.94	0.73
augMW logit BASE	−0.0028	0.086	0.086	0.084	0.94	0.77
augMW logit SIG	−0.0040	0.085	0.085	0.081	0.95	0.71
			$n = 100$			
MW	−0.0007	0.067	0.067	0.067	0.94	1.00
augMW probit BASE	−0.0005	0.059	0.059	0.059	0.95	0.78
augMW probit SIG	−0.0004	0.058	0.058	0.058	0.94	0.75
augMW logit BASE	−0.0005	0.059	0.059	0.059	0.95	0.78
augMW logit SIG	−0.0005	0.058	0.058	0.059	0.94	0.75
			$n = 200$			
MW	−0.0021	0.049	0.049	0.047	0.94	1.00
augMW probit BASE	−0.0013	0.043	0.043	0.043	0.94	0.84
augMW probit SIG	−0.0011	0.043	0.043	0.043	0.94	0.84
augMW logit BASE	−0.0013	0.043	0.043	0.043	0.94	0.84
augMW logit SIG	−0.0011	0.043	0.043	0.043	0.94	0.84

NOTE: BIAS, Monte Carlo Bias; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; AVESE, average of sandwich standard errors; COV, Monte Carlo coverage of 95% Wald confidence intervals; RE, relative efficiency is the Monte Carlo variance of the indicated estimator divided by the Monte Carlo variance of $U/(N_0N_1)$; MW, unadjusted estimator based on the Mann-Whitney U test statistic; augMW: augmented MW; probit, working PIM fitted using probit link; logit, working PIM fitted using logit link; BASE, adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model.

PIMs, both tables only show results for the logit link. Results for the probit link are shown in Table 6.9. Finally, Table 6.10 and Table 6.11 in Appendix 6.E show the percentage of times of all 1000 Monte Carlo replications that each variable is included in the working PIMs for both the probit and logit link function. They show that the baseline CD4 count (X_1) is added to almost all working PIMs (roughly in

Table 6.7: *Simulation results for tests based on* 1000 *Monte Carlo replications* ($n = 50$).

TEST	TYPE I	RE	POW	RE	TYPE I	RE	POW	RE
				$n = 50$				
	$\pi = 0.5$				$\pi = 0.75$			
Tests based on asymptotic approximations								
t-test	4.6	–	18.2	–	4.0	–	16.7	–
LR BASE	3.3	–	23.9	–	4.3	–	15.6	–
LR SIG	3.5	–	23.3	–	4.7	–	15.1	–
PIM logit BASE	6.6	–	34.4	–	7.5	–	32.8	–
PIM logit SIG	7.0	–	34.0	–	7.9	–	32.2	–
MW	5.3	–	25.3	–	6.2	–	24.7	–
augMW logit BASE	5.3	–	33.5	–	6.0	–	30.8	–
augMW logit SIG	5.4	–	32.0	–	7.0	–	29.6	–
ZHANG unadj	4.6	–	22.8	–	4.7	–	17.8	–
ZHANG BASE	4.8	–	32.8	–	6.9	–	30.4	–
ZHANG SIG	6.7	–	33.9	–	10.7	–	35.1	–
Permutation tests								
MW	4.3	1.00	22.5	1.00	4.1	1.00	17.9	1.00
augMW logit0 BASE	4.5	0.72	31.2	0.74	4.2	0.78	23.6	0.79
augMW logit0 SIG	4.5	0.66	28.4	0.68	4.1	0.74	22.1	0.77
augMW logit1 BASE	4.7	0.75	31.3	0.74	4.2	0.82	22.8	0.83
augMW logit1 SIG	4.4	0.70	28.1	0.72	4.5	0.80	21.2	0.83

NOTE: TYPE I: (unconditional) Type I error; RE: empirical variance of the augmented test statistic divided by the empirical variance of $U/(N_1N_0)$; POW: power; BASE: adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model; t-test: two-sample t-test; LR: linear regression; PIM logit: Wald test of $\tau_A = 0$ in PIM with logit link; MW: Mann-Whitney U test; augMW: augmented Mann-Withney U test; logit0: working PIM fitted excluding treatment using logit link; logit1: working PIM fitted including treatment using logit link; ZHANG: hypothesis tests described in Zhang et al. (2008), Section 5.

90% to 100% of all simulation experiments). Table 6.12 in Appendix 6.E shows this for the linear regression models (LR) and Table 6.13 and Table 6.14 in Appendix 6.E show this for the working models in the tests of Zhang et al. (2008), Section 5.

Both Table 6.7 and 6.8 show that the permutation based augmented Mann-

Table 6.8: *Simulation results for tests based on* 1000 *Monte Carlo replications* ($n = 100$).

TEST	TYPE I	RE	POW	RE	TYPE I	RE	POW	RE
				$n = 100$				
	$\pi = 0.5$				$\pi = 0.75$			
Tests based on asymptotic approximations								
t-test	4.5	–	37.9	–	4.2	–	31.6	–
LR BASE	4.7	–	45.2	–	4.2	–	28.3	–
LR SIG	4.7	–	44.0	–	4.6	–	28.0	–
PIM logit BASE	5.6	–	57.8	–	5.5	–	49.8	–
PIM logit SIG	5.9	–	58.0	–	6.0	–	49.0	–
MW	4.9	–	46.5	–	5.6	–	39.4	–
augMW logit BASE	4.6	–	57.3	–	5.5	–	49.4	–
augMW logit SIG	5.1	–	57.8	–	5.7	–	48.7	–
ZHANG unadj	4.7	–	45.6	–	4.9	–	36.6	–
ZHANG BASE	4.7	–	57.3	–	5.7	–	50.7	–
ZHANG SIG	5.2	–	59.3	–	7.7	–	50.8	–
Permutation tests								
MW	4.7	1.00	45.3	1.00	4.4	1.00	34.8	1.00
augMW logit0 BASE	4.2	0.78	56.2	0.79	4.1	0.77	43.1	0.74
augMW logit0 SIG	4.3	0.74	56.0	0.76	4.7	0.73	42.0	0.72
augMW logit1 BASE	4.2	0.77	56.1	0.79	4.2	0.78	43.5	0.74
augMW logit1 SIG	4.1	0.77	56.5	0.79	4.7	0.75	42.0	0.74

NOTE: TYPE I: (unconditional) Type I error; RE: empirical variance of the augmented test statistic divided by the empirical variance of $U/(N_1N_0)$; POW: power; BASE: adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model; t-test: two-sample t-test; LR: linear regression; PIM logit: Wald test of $\tau_A = 0$ in PIM with logit link; MW: Mann-Whitney U test; augMW: augmented Mann-Withney U test; logit0: working PIM fitted excluding treatment using logit link; logit1: working PIM fitted including treatment using logit link; ZHANG: hypothesis tests described in Zhang et al. (2008), Section 5.

Whitney U tests, the two-sample t-test and the classical Mann-Whitney U test attain the unconditional nominal level of 5%, as predicted by Theorem 6.5 (although they are sometimes slightly conservative). Although the tests for $\beta_A = 0$ in the linear regression models mostly attain the 5% level, they are somewhat conservative in the

Table 6.9: *Simulation results for tests using probit working models based on* 1000 *Monte Carlo replications.*

TEST	TYPE I	RE	POW	RE	TYPE I	RE	POW	RE
				$n = 50$				
		$\pi = 0.5$				$\pi = 0.75$		
Tests based on asymptotic approximations								
PIM probit BASE	6.6	–	35.4	–	8.1	–	34.0	–
PIM probit SIG	7.5	–	34.9	–	8.3	–	33.1	–
augMW probit BASE	5.4	–	33.5	–	5.9	–	30.7	–
augMW probit SIG	5.7	–	32.0	–	6.9	–	29.1	–
Permutation tests								
augMW probit0 BASE	4.6	0.74	31.0	0.74	4.1	0.74	23.5	0.74
augMW probit0 SIG	4.5	0.72	28.1	0.72	4.1	0.72	21.9	0.72
augMW probit1 BASE	4.7	0.76	31.1	0.74	3.9	0.78	22.8	0.83
augMW probit1 SIG	4.4	0.69	27.9	0.72	4.5	0.73	21.3	0.81
				$n = 100$				
		$\pi = 0.5$				$\pi = 0.75$		
Tests based on asymptotic approximations								
PIM probit BASE	5.7	–	58.1	–	5.9	–	50.3	–
PIM probit SIG	6.0	–	58.1	–	6.1	–	49.5	–
augMW probit BASE	4.5	–	57.5	–	5.5	–	49.2	–
augMW probit SIG	5.1	–	58.0	–	5.5	–	48.7	–
Permutation tests								
augMW probit0 BASE	4.2	0.74	56.2	0.74	4.1	0.74	43.2	0.74
augMW probit0 SIG	4.4	0.72	55.9	0.72	4.8	0.72	42.0	0.72
augMW probit1 BASE	4.2	0.75	56.2	0.79	4.1	0.78	43.1	0.76
augMW probit1 SIG	4.3	0.72	56.0	0.79	4.7	0.75	42.0	0.76

NOTE: TYPE I: (unconditional) Type I error; RE: empirical variance of the augmented test statistic divided by the empirical variance of $U/(N_1N_0)$; POW: power; BASE: adjustment for baseline CD4; SIG: adjustment for significant covariates in a univariate model; PIM probit: Wald test of $\tau_A = 0$ in PIM with probit link; augMW: augmented Mann-Withney U test; probit0: working PIM fitted excluding treatment using probit link; probit1: working PIM fitted including treatment using probit link.

balanced setting $\pi = 0.5$ for $n = 50$. Tests based on the null hypothesis $\tau_A = 0$ in a model for the CPI all show inflated Type I errors because of the poor approximation of the sandwich estimator for the standard error of the corresponding parameter

estimate $\hat{\tau}_{n,A}$ in small samples. This behavior is even worse in the unbalanced case with $\pi = 0.75$ and in combination with variable selection. Inflation of Type I errors is less severe at larger sample size. This may be due to the fact that the estimator of τ_A can be badly biased in finite samples. The Wald tests based on the asymptotic distribution of the test statistic $Z = (\hat{v}_{n,\text{adap}} - 0.5)/\{\widehat{\text{var}}_n(\hat{v}_{n,\text{adap})}\}^{1/2}$ for the estimators considered in the previous section also suffer from a slight inflation of Type I errors, especially in the unbalanced case and when combined with variable selection. A similar result is seen for the adjusted ZHANG tests, where inflation of Type I errors is more severe. The latter is somewhat unexpected because we would expect the tests of Zhang et al. (2008), which do not acknowledge estimation of the randomization probabilities, to be conservative.

Comparing the relative efficiency of the permutation based augmented Mann-Whitney U statistics with that of the ordinary permutation based Mann-Whitney U statistic, we observe important efficiency improvements by incorporating covariate information using the working PIM for the CPI fitted under the null, even though this working PIM is not necessarily correctly specified (the true CPI is unknown). For the different settings, we observe a decrease in variance of the test statistic ranging from roughly 20% to 35%. This is reflected in a gain in power to detect the alternative $P(Y \preceq Y^*|A = 0, A^* = 1) \neq 0.5$ as compared to the classical Mann-Whitney U test. This is most pronounced in the balanced ($\pi = 0.5$) case. Furthermore, it is observed that including treatment to the augmentation part does not yield additional power, thus confirming the adequacy of the computationally more efficient strategy outlined in Section 6.6. A possible reason for this might be that because including treatment to the augmentation part demands recalculating $\hat{\tau}_{n,g}$ for every permutation g, so a potential increase in power might be masked by a possible increase in variance by re-estimating $\hat{\tau}_{n,g}$ for every permutation g in small samples. The power for all augmented tests is larger than that for the two-sample t-test and the tests based on covariate adjustment via linear regression, which is due to the outlying outcomes. The power of the tests based on the null hypothesis $\tau_A = 0$ in a model for the CPI is often higher but not comparable because of inflated Type I errors. Finally, the power of the adjusted ZHANG tests is mostly comparable with that of the Wald tests presented in Section 6.5, although sometimes higher due to the more severe inflation of Type I errors for the ZHANG tests.

We conclude that the power of the classical Mann-Whitney U test can be drastically increased by using auxiliary covariate information via a PIM for the covariate-outcome association. The results suggest that it is sufficient to use only those covariates with highest predictive power for the outcome such as a baseline outcome measure. Interestingly, variable selection does not inflate the unconditional Type I error as predicted by the theory in Section 6.6.

6.9 Discussion

In this chapter, we have presented a robust adaptation of the Mann-Whitney U test, which may be more powerful than the classical Mann-Whitney U test, by allowing for covariate adjustment in randomized experiments. In the spirit of the work by Tsiatis et al. (2008), Zhang et al. (2008) and Moore and van der Laan (2009), we appealed to the theory of semiparametrics, from which we identified the class of all consistent and asymptotically normal estimators for the MPI under the model $\mathcal{M}_{\text{indep}}$ and characterized the most efficient one. We presented a locally efficient and adaptive estimation procedure for the MPI using so-called PIMs (Thas et al. 2012), which allows for covariate adjustment, without inducing bias under model misspecification. A sandwich estimator was also presented for large sample inference.

The semiparametric theory results presented here are related to the general semiparametric theory framework outlined in Schisterman and Rotnitzky (2001) for the estimation of the mean of K-sample U statistics with missingness in the outcome, explainable by auxiliary information. One crucial difference is that we allow for random treatment assignment and hence do not view the Mann-Whitney U test as a two-sample U-statistic with a kernel of degree one as presented in that paper, but as a one-sample U-statistic with a kernel of degree two. Nevertheless, we can still view our set-up as a two-sample problem where both samples are comprised of all subjects in the trial. In one sample, the observed data for individual i are given by $\boldsymbol{O}_i^{(1)} = (A_iY_i, A_i, \boldsymbol{X}_i)$, so $\boldsymbol{X}_i$ is always observed but the outcome is only observed if $A_i = 1$ and missing if $A_i = 0$. In the other sample, the observed data for individual i are given by $\boldsymbol{O}_i^{(0)}((1-A_i)Y_i, A_i, \boldsymbol{X}_i)$, so $\boldsymbol{X}_i$ is again always observed but

the outcome is now only observed if $A_i = 0$ and missing if $A_i = 1$. This set-up is different than in Schisterman and Rotnitzky (2001) where the different samples are assumed to be independent, which is clearly not the case here, leading to a different efficient estimator for the target parameter.

Because the Mann-Whitney test is often indicated in small sample settings, a permutation test was constructed based on the locally efficient estimator for the MPI as a test statistic. Using a test statistic that is guaranteed to be asymptotically less variable than the standardized Mann-Whitney U statistic U/N_0N_1, we anticipated the resulting test to be more powerful than the classical Mann-Whitney U test. This was confirmed in simulation studies and an application to an HIV clinical trial. Although we allow for random treatment group sizes, the permutation test remains valid for designs with fixed treatment group sizes. However, it remains to be seen if the estimator $\hat{\nu}_n(H_{\text{eff}})$ is efficient in that case. An attractive feature of our proposal is that, in contrast to regression adjustment via PIMs and the augmented tests based on asymptotic approximations, it preserves the Type I error under the null hypothesis (6.19) in small sample settings, even when combined with variable selection; this coincides with the results found in Stephens et al. (2013) for additive treatment effect measures. This is an attractive property, since it allows for data-adaptive variable selection in the construction of the working PIM without risking inflated Type I errors. We conjecture that this ability to obtain an accurate fit for the CPI makes it more likely to obtain a good approximation of the truth leading to enhanced efficiency for the MPI.

Caution is needed with the precise formulation of the null hypothesis of the permutation test in Section 6.6. The null hypothesis (6.19) is less stringent than the strong null hypothesis considered in the covariate-adjusted permutation tests of Rosenbaum (1984, 2002). This strong null hypothesis considers no treatment effect on any individual's outcome (that is, $Y_i(0) = Y_i(1)$ for all $i = 1,\ldots,n$). The null hypothesis (6.19) is stronger than the weak null hypothesis $P(Y \preceq Y^*|A = 0, A^* = 1) = 0.5$ in the sense that the MPI may equal one half even when the joint distribution of the observables $(Y, \boldsymbol{X})$ is different among treatment groups (see Chung and Romano (2011) for examples). Although an asymptotic Wald test based on $\hat{\nu}_{n,\text{adap}}$ will control the Type I error rate when this weak null hypothesis holds, even when the stronger null hypothesis (6.19) fails, this need not be the case for

the permutation test. The reason for this is that under the weak null hypothesis, the permutation null distribution of $n^{1/2}\widehat{T}_n(\boldsymbol{A},\boldsymbol{Y},\mathbb{X})$ will not necessarily converge to its unconditional distribution. Chung and Romano (2011, 2013) show that a permutation test based on the Mann-Whitney statistic which is appropriately studentized by a consistent estimator for its unconditional variance, achieves the correct asymptotic Type I error under the weak null hypothesis $P(Y \preceq Y^*|A = 0,A^* = 1) = 0.5$ (or more general the null hypothesis $P(Y \preceq Y^*|A = 0,A^* = 1) = \nu^*$ for any $\nu^* \in (0,1)$) but remains exact under the stronger null hypothesis of identical distributions of the outcome in both treatment groups. Further research is needed to investigate if these results still apply to the test statistic considered in Section 6.6 when studentization is based on the sandwich estimator (6.18). The sandwich estimator (6.18) should however be calculated using the working model m_0 (used in the construction of $\widehat{T}_n(\boldsymbol{A},\boldsymbol{Y},\mathbb{X})$) instead of the working model m.

As noted by a referee, given that a level α permutation test of the null hypothesis $P(Y \preceq Y^*|A = 0,A^* = 1) = \nu^*$ for any $\nu^* \in (0,1)$ can be constructed, one could in principle construct an exact $(1-\alpha)100\%$ confidence interval for the MPI based on inverting permutation tests. However, it is unclear how permutations of the treatment assignment can be performed in a way such that the hypothesis $P(Y \preceq Y^*|A = 0,A^* = 1) = \nu^*$ for a $\nu^* \neq 0.5$ is preserved. Lehmann (1963) shows that an exact $(1-\alpha)100\%$ confidence interval can be obtained from a Mann-Whitney U test for a shift parameter if one is additionally willing to assume a location shift model. However, without making any additional distributional assumptions, this will not lead to an exact $(1-\alpha)100\%$ confidence interval for the MPI. Also building on parametric assumptions, Newcombe (2006) constructs an exact $(1-\alpha)100\%$ confidence interval for the MPI using a tail area modeling approach rather than inverting a permutation test. This demands knowledge of the probabilities of each possible sequence of the ordered outcomes for treated and untreated subjects. Unfortunately, to our knowledge, it is unknown in the absence of distributional assumptions how to permute the vector of treatment assignments so that the hypothesis that the MPI equals ν^* for $\nu^* \neq 0.5$ is maintained.

An efficiency benefit of the augmented Mann-Whitney U test statistic relative to the unadjusted Mann-Whitney U test statistic is only guaranteed when the working model for the PIM is correctly specified. Misspecification of the PIM does not

alter consistency but can affect the asymptotic variance. However, as illustrated in the data analysis and the simulation studies, we did observe an efficiency benefit, even under PIM misspecification. Guaranteed efficiency improvement, even under misspecification of the PIM, could potentially be obtained by estimating the parameters indexing the PIM as those minimizing the asymptotic variance of the estimator of the MPI, along the lines of the empirical efficiency maximization procedure (Rubin and van der Laan 2008; Cao et al. 2009). Although it could additionally be of interest to extend the Targeted Maximum Likelihood Estimator (TMLE, van der Laan and Rubin (2006)) to the locally efficient estimator of the MPI, we did not consider this because (6.7) is already a substitution estimator like TMLE and because we are not aware of learning algorithms for the CPI.

A limitation of our proposal is that covariate adjustment may lead to a severe decrease in sample size when there is substantial missingness in the covariates and a complete-case analysis is performed. An efficiency benefit as compared to an unadjusted analysis may then be lost. In such case, we recommend using multiple imputation based on an imputation model that only includes covariates but no outcome and exposure. The reason for this is that in this manner, we do not induce bias in the estimator for the MPI, even when the imputation model is misspecified and regardless of the missing data mechanism.

Finally, the semiparametric results presented here can be generalized to adjust for confounding in observational studies when interest lies in the MPI $P\{Y(0) \preceq Y^*(1)\}$ where $Y(a)$ denotes the counterfactual outcome for treatment level $a \in \{0,1\}$. Instead of estimating the randomization probability π using the proportion of treated individuals, a propensity score model for the probability of being treated given confounders is now needed. Adjustment for confounding based on adaptations of the Mann-Whitney U test have already been suggested in Wu et al. (2013); Chen et al. (2013) but they lack general semiparametric efficiency results. We will report on this extension in the next chapter.

6.A Estimating Equations and Asymptotic Theory for PIMs

A detailed discussion on how to estimate and do inference on the parameters indexing PIMs can be found in Thas et al. (2012). Below we briefly discuss the most important results. Consider a random sample $\{\boldsymbol{O}_i = (Y_i, \boldsymbol{X}_i) : i = 1, \ldots, n\}$, an outcome and a vector of covariates. Instead of modeling the conditional mean outcome, a PIM models the conditional probabilistic index (CPI): $P(Y_i \preceq Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j) = P(Y_i < Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j) + 0.5P(Y_i = Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j)$. The class of models considered in Thas et al. (2012) is defined as

$$P(Y_i \preceq Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau}) = m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau}) = g^{-1}(\boldsymbol{\tau}^T \boldsymbol{Z}_{ij}), \quad (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n, \qquad (6.21)$$

with $\mathcal{X}_n$ the set of pairs of predictors $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ for which the model is defined. The function m has range $[0,1]$ and $\boldsymbol{\tau}$ is a p-dimensional parameter and m should satisfy $m(\boldsymbol{X}_i, \boldsymbol{X}_i; \boldsymbol{\tau}) = 0.5$ and $m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau}) + m(\boldsymbol{X}_j, \boldsymbol{X}_i; \boldsymbol{\tau}) = 1$ for $(\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n$ and $(\boldsymbol{X}_j, \boldsymbol{X}_i) \in \mathcal{X}_n$. Similar as for generalized linear models, $m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\tau})$ is restricted to be in the class $g^{-1}(\boldsymbol{\tau}^T \boldsymbol{Z}_{ij})$, with g an appropriate link function and $\boldsymbol{Z}_{ij}$ is a function of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. Besides some technical conditions, (6.21) is the sole restriction the PIM makes about the conditional distribution of Y given $\boldsymbol{X}$ and hence constitutes a semiparametric model. A meaningful choice in many applications is $\boldsymbol{Z}_{ij} = \boldsymbol{X}_j - \boldsymbol{X}_i$ and since the PIM models a probability, convenient link functions are the logit link and probit link functions but sometimes the identity link may be convenient. To obtain a consistent and asymptotically normally distributed estimator for $\boldsymbol{\tau}$ and a consistent estimator for its asymptotic variance, define the **pseudo-observations** $I_{ij} = I(Y_i \preceq Y_j)$ as $I(Y_i < Y_j) + 0.5I(Y_i = Y_j)$. Let $\mathcal{I}_n$ denote the set of indices (i, j) such that $(\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n$. A consistent estimator $\hat{\boldsymbol{\tau}}_n$ of $\boldsymbol{\tau}$ can then be found as the solution to the estimating equations

$$\boldsymbol{U}_n(\boldsymbol{\tau}) = \sum_{(i,j) \in \mathcal{I}_n} \boldsymbol{U}_{ij}(\boldsymbol{\tau}) = \sum_{(i,j) \in \mathcal{I}_n} \boldsymbol{A}(\boldsymbol{Z}_{ij}; \boldsymbol{\tau}) \{I_{ij} - g^{-1}(\boldsymbol{\tau}^T \boldsymbol{Z}_{ij})\} = \boldsymbol{0}, \qquad (6.22)$$

with $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\tau}) = \{\partial g^{-1}(\boldsymbol{\tau}^T\boldsymbol{Z}_{ij})/\partial\boldsymbol{\tau}\}\boldsymbol{V}^{-1}\{g^{-1}(\boldsymbol{\tau}^T\boldsymbol{Z}_{ij})\}$ where $\boldsymbol{V}\{g^{-1}(\boldsymbol{\tau}^T\boldsymbol{Z}_{ij})\} = \text{var}(I_{ij}|\boldsymbol{Z}_{ij})$. Thas et al. (2012) show that $\hat{\boldsymbol{\tau}}_n$ is asymptotically normal with mean $\boldsymbol{\tau}$ and covariance matrix $\Sigma(\boldsymbol{\tau})$ which can be consistently estimated using the sandwich estimator

$$\hat{\Sigma}_n(\hat{\boldsymbol{\tau}}_n) = \left\{\sum_{(i,j)\in\mathcal{I}_n}\frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\tau}}_n)}{\partial\boldsymbol{\tau}^T}\right\}^{-1}\left\{\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n}\phi_{ijkl}\boldsymbol{U}_{ij}(\hat{\boldsymbol{\tau}}_n)\boldsymbol{U}_{kl}^T(\hat{\boldsymbol{\tau}}_n)\right\} \times\left\{\sum_{(i,j)\in\mathcal{I}_n}\frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\tau}}_n)}{\partial\boldsymbol{\tau}}\right\}^{-1}, \tag{6.23}$$

where ϕ_{ijkl} equals 1 if I_{ij} and I_{kl} are correlated and 0 otherwise.

6.B Equivalence of the Mann-Whitney Test and a PIM with Logit Link

Below we illustrate that the Mann-Whitney U test can also be obtained using a logit link function. That is, we posit the model $\text{logit}P(Y \preceq Y^*|A,A^*;\tau) = \tau(A^* - A)$. The estimating equation (6.22) becomes

$$\sum_{i=1}^{n}\sum_{j\neq i}(A_j - A_i)[I_{ij} - \text{expit}\{\tau(A_j - A_i)\}] = 0$$

because $\partial\text{expit}\{\tau(A_j - A_i)\}/\partial\tau = (A_j - A_i)\text{expit}\{\tau(A_j - A_i)\}[1 - \text{expit}\{\tau(A_j - A_i)\}]$ and the conditional variance

$$\text{var}(I_{ij}|A_i,A_j) = \text{expit}\{\tau(A_j - A_i)\}[1 - \text{expit}\{\tau(A_j - A_i)\}].$$

Rewriting $(A_j - A_i)$ as $A_j(1 - A_i) - A_i(1 - A_j)$ gives

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\sum_{j\neq i}A_j(1 - A_i)\{I_{ij} - \text{expit}(\tau)\} - \sum_{i=1}^{n}\sum_{j\neq i}A_i(1 - A_j)\{I_{ij} - \text{expit}(-\tau)\} \\ &= \sum_{i=1}^{n}\sum_{j\neq i}A_j(1 - A_i)\{I_{ij} - \text{expit}(\tau)\} + \sum_{i=1}^{n}\sum_{j\neq i}A_i(1 - A_j)\{I_{ji} - \text{expit}(\tau)\} \end{aligned}$$

because $I_{ij} = 1 - I_{ji}$ and $\text{expit}(-\tau) = 1 - \text{expit}(\tau)$. Interchanging i and j in the second summation of the latter equation gives

$$0 = \sum_{i=1}^{n} \sum_{j \neq i} A_j (1 - A_i) \{ I_{ij} - \text{expit}(\tau) \}.$$

Solving for τ yields the estimator

$$\begin{aligned} \hat{\tau}_n &= \text{logit} \left\{ \sum_{i=1}^{n} \sum_{j \neq i} A_j (1 - A_i) I_{ij} / \sum_{i=1}^{n} \sum_{j \neq i} A_j (1 - A_i) \right\} \\ &= \text{logit}(U / N_0 N_1) \end{aligned}$$

This shows that $P(Y \preceq Y^* | A = 0, A^* = 1; \hat{\tau}_n) = U / N_0 N_1$.

6.C Connection to *the improved hypothesis tests of Zhang et al. (2008)*

Below, we elaborate on the mathematical details connecting the semiparametric efficiency results as presented in Section 6.5 and the results from Zhang et al. (2008). Specifically, we show that the augmented test statistic of Zhang et al. (2008), see their equation (18) and (19), is proportional to the Hájek projection of the augmented test statistic considered in this chapter, but only when the unadjusted test statistic is calculated under the null hypothesis.

The approach of Zhang et al. (2008) proceeds by augmenting a test statistic that is asymptotically equivalent to the Hájek projection $\ell(Y, A)$ of the Mann-Whitney U test statistic, calculated under the null hypothesis H_0 that the outcome Y is independent of the treatment assignment A:

$$\ell(Y, A) = (A - \pi) \{ S(Y) - 0.5 \}, \tag{6.24}$$

with $S(y) = 1 - P(Y \leq y)$ the survival function of the outcome Y. Now consider the Hájek projection of the Mann-Whitney U test statistic considered in equation (6.9). It is slightly different from the results in van der Vaart (1998), sec. 12.2, by

not fixing the treatment groups and viewing the Mann-Whitney U test statistic as a one-sample U-statistic with kernel of degree 2 rather than a two-sample U-statistic with kernel of degree 1. Under the null hypothesis that A is independent of Y, or equivalently, that those treated have the same survival function as those not treated, (6.9) simplifies to a form which is comparable with the test statistic considered in Zhang et al. (2008). We have that under H_0, $a_1(Y_i) = E\{(A_j/\pi)I_{ij}|Y_i\} = S(Y_i)$ and $a_2(Y_i) = E\{(1-A_j)/(1-\pi)I_{ji}|Y_i\} = 1-S(Y_i)$ and $\nu_0 = 0.5$ such that

$$\begin{aligned}\tilde{\nu}_{n,0}^{(H_0)} &= -n^{-1}\sum_{i=1}^{n}\frac{1}{\pi(1-\pi)}\{\ell(Y_i,A_i)-(A_i-\pi)0.5(1-2\pi)\}\\ &= -\frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}\ell(Y_i,A_i)+\frac{0.5-\pi}{\pi(1-\pi)}n^{-1}\sum_{i=1}^{n}(A_i-\pi)\\ &= -\frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}\ell(Y_i,A_i)+o_p(1),\end{aligned}$$

6

showing that, under H_0, (6.9) is proportional and asymptotically equivalent to the test statistic considered in Zhang et al. (2008), resulting in asymptotically equivalent standardized test statistics (under H_0).

In Zhang et al. (2008), $\ell(Y,A)$ is optimally augmented (in the sense is has lowest variance among a certain class of functions) to obtain $n^{-1}\sum_{i=1}^{n}\ell^*(Y_i,\boldsymbol{X}_i,A_i)$, with

$$\begin{aligned}&\ell^*(Y_i,\boldsymbol{X}_i,A_i)\\ &\quad=\ell(Y_i,A_i)+(A_i-\pi)\left[E\{\ell(Y_i,0)|A_i=0,\boldsymbol{X}_i\}-E\{\ell(Y_i,1)|A_i=1,\boldsymbol{X}_i\}\right].\end{aligned}$$

In contrast, the Hájek projection of (6.9) is given by

$$\begin{aligned}\tilde{\nu}_{n,0}^{*} = n^{-1}\sum_{i=1}^{n}\Bigg(&\frac{1-A_i}{1-\pi}a_1(Y_i)-\nu_0+\frac{A_i}{\pi}a_2(Y_i)-\nu_0\\ &+(A_i-\pi)\left[\frac{E\{a_1(Y_i)|A_i=0,\boldsymbol{X}_i\}}{1-\pi}-\frac{E\{a_2(Y_i)|A_i=1,\boldsymbol{X}_i\}}{\pi}\right]\Bigg).\end{aligned}$$

However, it turns out that when the functions a_1 and a_2 are calculated under H_0 and ν_0 is taken to be 0.5, this Hájek projection is identically (up to a constant multiple) as $n^{-1}\sum_{i=1}^{n}\ell^*(Y_i,\boldsymbol{X}_i,A_i)$. Indeed, using the fact that $a_1(Y_i) = S(Y_i)$ and

$a_2(Y_i) = 1 - S(Y_i)$ under H_0, we find that

$$\begin{aligned}
\tilde{v}_{n,0}^{*,(H_0)} =& -\frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}\ell(Y_i,A_i) + \frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}(A_i-\pi)\times[(0.5-\pi) \\
&+\pi E\{S(Y_i)|A_i=0,\boldsymbol{X}_i\} + (1-\pi)E\{S(Y_i)|A_i=1,\boldsymbol{X}_i\} - (1-\pi)] \\
=& -\frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}\ell(Y_i,A_i) - \frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}(A_i-\pi) \\
&\times[-\pi E\{S(Y_i)-0.5|A_i=0,\boldsymbol{X}_i\} - (1-\pi)E\{S(Y_i)-0.5|A_i=1,\boldsymbol{X}_i\}] \\
=& -\frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}\ell(Y_i,A_i) - \frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}(A_i-\pi) \\
&\times[E\{\ell(Y_i,0)|A_i=0,\boldsymbol{X}_i\} - E\{\ell(Y_i,1)|A_i=1,\boldsymbol{X}_i\}] \\
=& -\frac{n^{-1}}{\pi(1-\pi)}\sum_{i=1}^{n}\ell^*(Y_i,\boldsymbol{X}_i,A_i).
\end{aligned}$$

This calculation shows that the augmented test statistic $n^{-1}\sum_{i=1}^{n}\ell^*(Y_i,\boldsymbol{X}_i,A_i)$ considered in Zhang et al. (2008) corresponds to (up to a constant which disappears after standardizing) the Hájek projection $\tilde{v}_{n,0}^{*,(H_0)}$ of (6.9), given that the unadjusted test statistic is calculated under H_0.

6.D R-Functions

In the two R-functions given below, the PIMs are fitted using the `pim` package for R available on R-forge (De Neve and Sabbe 2013).

6.D.1 Estimation and asymptotic inference

Below, we provide an R-function that implements the locally efficient estimation procedure for the MPI from Section 6.5.2. As input, the function uses the outcome `y`, the treatment indicator `a`, the auxiliary covariates `x`, the link function `link=c("probit","logit")` for the PIM and the level of significance `alpha`. As output, the function delivers the estimate `est` which equals $\hat{v}_{n,\text{adap}}$, the standard error `se` which equals $\{\widehat{\text{var}}_n(\hat{v}_{n,\text{adap}})\}^{1/2}$ based on equation (6.18), a $(1-\alpha)100\%$ Wald confidence interval `CI`, the Wald statistic `Wald.statistic`

and the corresponding p-value `p.value`. We also provide an example where the locally efficient estimation procedure is applied to a random subsample of the ACTG 175 clinical trial with adjustment for baseline CD4 count (X_1).

R-function

```
augmented.MW<-function(y,a,x,link=c("probit","logit"),alpha){
  n<-length(a)
  p<-sum(a)/n
  data.sub<-data.frame(cbind(y,a,x))
  x.dat<-as.data.frame(x)
  x<-as.matrix(x)

  # Point estimate:
  pim.fit<-pim(formula=as.formula(paste("y~a+",
               paste(names(x.dat),collapse="+"))),
               link=link,data=data.sub)
  coef<-pim.fit$coef
  MW<-1-wilcox.test(y~a,exact=FALSE)$statistic/
            (sum(a)*sum(1-a))
  {
     if(link=="probit"){
        augMW<-MW
        for(i in 1:n){augMW<-augMW+sum((1/(n*(n-1))-
           (1-a[i])*a[-i]/(sum(a)*(n-sum(a))))*pnorm(coef[1]+
           t(-x[i,]+t(x[-i,]))%*%coef[2:(dim(x)[2]+1)]))}
     }else if(link=="logit"){
        augMW<-0
        for(i in 1:n){augMW<-augMW+sum(expit(coef[1]+
           t(-x[i,]+t(x[-i,]))%*%coef[2:(dim(x)[2]+1)]))/
           (n*(n-1))}
    }else{print("not a valid link function")}
  }

  # Standard error
  pseudo.y<-pseudo(y)
  a1.hat<-sapply(A=a,pseudo.Y=pseudo.y,1:length(a),vec.a1hat)
  a2.hat<-sapply(A=a,pseudo.Y=pseudo.y,1:length(a),vec.a2hat)
  phi0<-vec.phi0(a,a1.hat,a2.hat,augMW)
  {
    if(link=="probit"){
      pred<-t(sapply(coef=coef,x=x,1:dim(x)[1],vec.pred.probit))
      alphahat<-sapply(A=a,pred=pred,1:length(a),vec.alphahat)
      phiest<-phi0+(a-p)*(alphahat-mean(alphahat)
         +mean((1-a)*a1.hat/(1-p)^2-a*a2.hat/p^2))
    }else if(link=="logit"){
```

```
      pred<-t(sapply(coef=coef,x=x,1:dim(x)[1],vec.pred.logit))
      alphahat<-sapply(A=a,pred=pred,1:length(a),vec.alphahat)
      phiest<-phi0+(a-p)*alphahat
    }else{print("not a valid link function")}
  }
  se.augMW<-sqrt(mean(phiest^2)/n)

  # 95% CI
  CI<-augMW+c(-1,1)*se.augMW*qnorm(1-alpha/2)

  # Wald test statistic:
  W<-(augMW-0.5)/se.augMW

  # p-value Wald test:
  p.value<-2*pnorm(abs(W),lower.tail=FALSE)

  return(list(est=augMW,se=se.augMW,CI=CI,Wald.statistic=W,
     p.value=p.value))
}
```

Auxiliary functions

```
expit<-function(x){exp(x)/(1+exp(x))}

pseudo<-function(Y){
   I<-matrix(rep(Y,length(Y)),ncol=length(Y),byrow=TRUE)
   I<-ifelse(I<Y,1,ifelse(I==Y,0.5,0))
   return(t(I))
}
vec.a1hat<-function(A,pseudo.Y,i){
   sum(A[-i]*pseudo.Y[i,-i])/((sum(A)/length(A))*(length(A)-1))
}
vec.a2hat<-function(A,pseudo.Y,i){
   sum((1-A[-i])*pseudo.Y[-i,i])/((sum(1-A)/length(A))*
   (length(A)-1))
}
vec.phi0<-function(A,a1.hat,a2.hat,est){
   p<-sum(A)/length(A)
   (1-A)/(1-p)*a1.hat+A/p*a2.hat-2*est
}
vec.pred.probit<-function(coef,x,i){
   pnorm(coef[1]+t(-x[i,]+t(x[,]))%*%coef[2:(dim(x)[2]+1)])
}
vec.pred.logit<-function(coef,x,i){
   expit(coef[1]+t(-x[i,]+t(x[,]))%*%coef[2:(dim(x)[2]+1)])
}
vec.alphahat<-function(A,pred,i){
```

```
   p<-sum(A)/length(A)
   sum(pred[i,-i]/(1-p)-pred[-i,i]/p)/(length(A)-1)
}
```

Example

```
library(speff2trial)
library(pim)
data(ACTG175)
attach(ACTG175)
A<-treat;Y<-cd420;
X1<-cd40;X2<-cd80;X3<-age;X4<-wtkg;
X5<-karnof;X6<-hemo;X7<-homo;X8<-drugs;
X9<-race;X10<-gender;X11<-str2;X12<-symptom;
data=data.frame(cbind(Y,A,X1,X2,X3,X4,X5,X6,
                      X7,X8,X9,X10,X11,X12))

set.seed(1)
sub<-sample.int(length(A),200,replace = FALSE)
data.sub<-data[sub,]
a<-data.sub$A;
y<-data.sub$Y;
x1<-data.sub$X1;x2<-data.sub$X2;x3<-data.sub$X3;
x4<-data.sub$X4;x5<-data.sub$X5;x6<-data.sub$X6;
x7<-data.sub$X7;x8<-data.sub$X8;x9<-data.sub$X9;
x10<-data.sub$X10;x11<-data.sub$X11;x12<-data.sub$X12;
x<-cbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12)

augmented.MW(y,a,x=x1,link="logit",alpha=0.05)
```

6.D.2 Permutation test

Below, we provide an R-function to perform the permutation testing procedure from Section 6.6.2. As input, the function uses the outcome `y`, the treatment indicator `a`, the auxiliary covariates `x`, the number of random permutations `perm` to be used, the link function `link=c("probit","logit")` for the PIM and the level of significance `alpha`. As output, the function delivers the observed test statistic `test.obs` which equals $\widehat{T}_n(\boldsymbol{A},\boldsymbol{Y},\mathbb{X})$, the p-value `pvalue`, the rejection boundaries `test.left` and `test.right` of the permutation null distribution, a rejection indicator `rej.test` (1 if rejected, 0 if not) and a vector containing the approximation of the permutation null distribution of the test statstic `test.star`.

We also provide an example where the augmented Mann-Whitney U test is applied to a random subsample of the ACTG 175 clinical trial with adjustment for baseline CD4 count.

R-function

```
perm.cov<-function(y,a,x,perm,link=c("probit","logit"),alpha){
  n<-length(a);p<-sum(a)/n
  data.sub<-data.frame(cbind(y,a,x))
  x.dat<-as.data.frame(x);x<-as.matrix(x)

  pim.fit<-pim(formula=as.formula(paste("y~",paste(names(x.dat),
     collapse = " + "))),link=link,data=data.sub)
  coef<-as.matrix(pim.fit$coef)

  MW<-1-wilcox.test(y~a,exact=FALSE)$statistic/(sum(a)*sum(1-a))

  # Observed test statistic
  test.obs<-MW
  {
     if(link=="probit"){
        for(i in 1:n){test.obs<-test.obs+sum((1/(n*(n-1))-
           (1-a[i])*a[-i]/(sum(a)*(n-sum(a))))*
           pnorm(t(-x[i,]+t(x[-i,]))%*%coef))}
     }else if(link=="logit"){
        for(i in 1:n){test.obs<-test.obs+sum((1/(n*(n-1))-
           (1-a[i])*a[-i]/(sum(a)*(n-sum(a))))*
           expit(t(-x[i,]+t(x[-i,]))%*%coef))}
     }else{print("not a valid link function")}
  }

  # Permutation null distribution
  test.star<-rep(NA,perm)
  for(j in 1:perm){test.star[j]<-perm.dist(y,a,x,j,link,coef)}
  pvalue<-2*min(mean(test.obs>=test.star),
                              mean(test.obs<=test.star))
  test.left<-quantile(test.star,alpha/2)
  test.right<-quantile(test.star,1-alpha/2)
  rej.test<-ifelse(test.obs>=test.right|test.obs<=test.left,1,0)
  return(list(test.obs=test.obs,pvalue=pvalue,
              test.left=test.left,test.right=test.right,
              rej.test=rej.test,test.star=test.star))
}
```

Auxiliary functions

```
expit<-function(x){exp(x)/(1+exp(x))}

perm.dist<-function(y,a,x,j,link,coef){
  n<-length(a)
  perm.a<-sample(a,replace=FALSE)
  STAR.MW<-1-wilcox.test(y~perm.a,exact=FALSE)$statistic/
     (sum(perm.a)*sum(1-perm.a))
  STAR<-STAR.MW
  {
     if(link=="probit"){
        for(i in 1:n){STAR<-STAR+sum((1/(n*(n-1))-
           (1-perm.a[i])*perm.a[-i]/(sum(perm.a)*
           (n-sum(perm.a))))*
           pnorm(t(-x[i,]+t(x[-i,]))%*%coef))}
     }else if(link=="logit"){
        for(i in 1:n){STAR<-STAR+sum((1/(n*(n-1))-
        (1-perm.a[i])*perm.a[-i]/(sum(perm.a)*
        (n-sum(perm.a))))*
        expit(t(-x[i,]+t(x[-i,]))%*%coef))}
     }else{print("not a valid link function")}
  }
  return(STAR)
}
```

Example

```
library(speff2trial)
library(pim)
data(ACTG175)
attach(ACTG175)
A<-treat;Y<-cd420;
X1<-cd40;X2<-cd80;X3<-age;X4<-wtkg;
X5<-karnof;X6<-hemo;X7<-homo;X8<-drugs;
X9<-race;X10<-gender;X11<-str2;X12<-symptom;
data=data.frame(cbind(Y,A,X1,X2,X3,X4,X5,X6,
                      X7,X8,X9,X10,X11,X12))

set.seed(1)
sub<-sample.int(length(A), 50, replace = FALSE)
data.sub<-data[sub,]
a<-data.sub$A;
y<-data.sub$Y;
x1<-data.sub$X1;x2<-data.sub$X2;x3<-data.sub$X3;
```

```
x4<-data.sub$X4;x5<-data.sub$X5;x6<-data.sub$X6;
x7<-data.sub$X7;x8<-data.sub$X8;x9<-data.sub$X9;
x10<-data.sub$X10;x11<-data.sub$X11;x12<-data.sub$X12;
x<-cbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12)

aug.test<-perm.cov(y,a,x=x1,perm=10000,link="logit",alpha=0.05)
aug.test$test.obs
aug.test$pvalue
aug.test$rej.test

hist(aug.test$test.star,main="Permutation Null Distribution",
   xlab="augmented MW test statistic")
abline(v=aug.test$test.left,lty=2,col="red",lwd=2)
abline(v=aug.test$test.right,lty=2,col="red",lwd=2)
abline(v=aug.test$test.obs,lty=1,col="blue",lwd=2)
```

6.E Results Variable Selection Simulation Studies

6.E.1 Results for the probabilistic index models (PIMs)

Table 6.10 and Table 6.11 show the percentage of times that each variable is included in the working PIMs for the probit and logit link for the variable selection procedure in the simulation studies.

Table 6.10: *Percentage variables included in the PIMs for the SIG strategy based on* 1000 *Monte Carlo replications when data is generated under* H_0.

COV	$n = 50$ $\pi = 0.5$	$n = 100$ $\pi = 0.5$	$n = 50$ $\pi = 0.75$	$n = 100$ $\pi = 0.75$	$n = 50$ $\pi = 0.5$	$n = 100$ $\pi = 0.5$	$n = 50$ $\pi = 0.75$	$n = 100$ $\pi = 0.75$
	probit PIM				logit PIM			
X_1	95.0	99.9	93.0	99.8	94.2	99.8	92.1	99.8
X_2	9.2	6.6	9.7	7.0	8.8	6.6	9.4	6.8
X_3	9.3	8.3	9.3	8.7	8.2	8.2	8.7	8.4
X_4	6.4	6.7	7.4	6.4	6.1	6.5	6.8	6.2
X_5	12.6	13.9	12.8	13.1	11.9	13.4	11.6	12.9
X_6	18.0	12.8	17.7	11.3	17.1	12.3	16.9	10.4
X_7	5.6	5.7	6.2	5.5	5.6	5.4	6.1	5.3
X_8	12.4	8.8	12.1	8.8	11.5	8.4	11.4	7.8
X_9	9.0	8.0	8.8	8.2	8.2	7.9	8.6	8.0
X_{10}	9.4	6.8	8.5	6.5	8.9	6.6	8.3	6.1
X_{11}	13.5	19.4	12.7	18.4	12.8	19.2	12.0	18.2
X_{12}	9.3	8.4	10.2	9.2	8.9	7.8	9.4	8.6

NOTE: COV, covariate; logit, working PIM fitted using logit link; probit, working PIM fitted using probit link.

Table 6.11: *Percentage variables included in the PIMs for the SIG strategy based on* 1000 *Monte Carlo replications when data is generated under the alternative.*

COV	$n=50$ $\pi=0.5$	$n=100$ $\pi=0.5$	$n=50$ $\pi=0.75$	$n=100$ $\pi=0.75$	$n=50$ $\pi=0.5$	$n=100$ $\pi=0.5$	$n=50$ $\pi=0.75$	$n=100$ $\pi=0.75$
	probit PIM				logit PIM			
X_1	93.3	100.0	92.0	99.6	92.9	100.0	91.2	99.6
X_2	10.0	6.3	8.9	7.6	9.4	6.1	8.4	7.1
X_3	8.2	7.7	9.2	8.6	7.9	7.5	8.9	8.4
X_4	6.8	5.9	8.2	6.3	6.2	5.7	7.5	6.0
X_5	11.8	13.8	12.3	13.2	11.3	13.0	11.5	12.8
X_6	17.3	11.8	16.6	12.1	17.0	11.5	15.5	11.3
X_7	6.6	4.9	5.3	5.0	6.3	4.6	5.2	4.7
X_8	13.1	8.0	11.8	9.1	12.1	7.4	11.0	8.9
X_9	8.3	8.3	9.3	8.2	8.2	8.2	9.2	7.9
X_{10}	9.9	6.3	9.6	7.3	9.4	6.1	9.3	6.8
X_{11}	13.6	20.7	2.6	18.1	12.9	20.3	11.8	17.9
X_{12}	8.4	7.2	10.6	8.6	7.9	7.1	10.0	8.3

NOTE: COV, covariate; logit, working PIM fitted using logit link; probit, working PIM fitted using probit link.

6.E.2 Results for the linear regression (LR)

Table 6.12 shows the percentage of times that each variable is included in the LR model for the variable selection procedure in the simulation studies.

Table 6.12: *Percentage variables included in the LR model for the SIG strategy based on* 1000 *Monte Carlo replications when data is generated under* H_0 *and the alternative.*

COV	$n=50$ $\pi=0.5$	$n=100$ $\pi=0.5$	$n=50$ $\pi=0.75$	$n=100$ $\pi=0.75$	$n=50$ $\pi=0.5$	$n=100$ $\pi=0.5$	$n=50$ $\pi=0.75$	$n=100$ $\pi=0.75$
	under H_0				under the alternative			
X_1	87.2	97.9	84.4	97.4	85.8	98.2	83.8	96.9
X_2	6.5	5.4	6.5	5.2	6.6	4.9	6.5	5.5
X_3	6.1	5.7	6.1	6.0	6.3	5.8	6.3	6.2
X_4	4.6	4.9	4.6	4.4	4.6	4.9	4.8	4.6
X_5	8.4	10.8	8.3	11.1	7.7	11.5	7.9	10.8
X_6	6.1	7.5	5.8	7.4	5.7	6.8	5.5	7.9
X_7	4.5	4.5	4.6	4.2	4.6	3.7	4.9	4.1
X_8	4.0	5.1	4.0	4.7	3.8	4.6	4.2	4.7
X_9	6.1	6.2	6.4	6.4	5.9	6.0	6.0	6.4
X_{10}	4.1	4.4	4.2	4.5	4.4	5.0	4.8	5.0
X_{11}	10.2	15.7	9.6	14.3	10.4	16.1	9.3	14.3
X_{12}	5.0	5.1	5.8	5.8	4.0	5.1	6.3	6.2

NOTE: COV, covariate.

6.E.3 Results for the improved hypothesis tests of ZHANG

Table 6.13 and Table 6.14 show the percentage of times that each variable is included in the working models used in the construction of the improved hypothesis tests of Zhang et al. (2008) outlined above for the variable selection procedure in the simulation studies.

Table 6.13: *Percentage variables included in the working models used in the construction of the improved hypothesis tests of ZHANG for the SIG strategy based on* 1000 *Monte Carlo replications when data is generated under* H_0.

COV	$n=50$ $\pi=0.5$	$n=100$ $\pi=0.5$	$n=50$ $\pi=0.75$	$n=100$ $\pi=0.75$	$n=50$ $\pi=0.5$	$n=100$ $\pi=0.5$	$n=50$ $\pi=0.75$	$n=100$ $\pi=0.75$
	ZHANG 0				ZHANG 1			
X_1	71.7	94.5	38.5	70.9	71.3	95.5	85.1	98.3
X_2	5.8	5.6	5.3	5.2	5.9	5.8	5.5	6.1
X_3	6.0	5.9	5.0	5.4	4.3	6.6	5.1	5.7
X_4	4.4	4.8	4.6	5.0	4.4	4.8	5.3	5.0
X_5	6.5	7.9	5.5	5.7	7.1	7.6	6.1	9.8
X_6	3.2	4.5	1.6	2.8	2.8	4.6	3.5	5.3
X_7	5.1	5.7	3.8	5.0	4.5	4.3	4.8	5.0
X_8	4.0	4.4	2.6	4.4	4.1	6.4	4.6	3.3
X_9	4.7	7.1	5.6	5.9	4.9	5.2	5.3	7.4
X_{10}	4.9	6.3	3.0	4.2	5.5	4.1	4.9	4.9
X_{11}	8.5	10.8	5.7	8.5	8.0	13.0	10.3	15.4
X_{12}	5.7	5.7	4.6	4.6	3.7	3.3	5.8	6.8

NOTE: COV, covariate; ZHANG 0, working model for the hypothesis test described in Zhang et al. (2008) for the $A=0$ group; ZHANG 1, working model for the hypothesis test described in Zhang et al. (2008) for the $A=1$ group.

Table 6.14: *Percentage variables included in the working models used in the construction of the improved hypothesis tests of ZHANG for the SIG strategy based on* 1000 *Monte Carlo replications when data is generated under the alternative.*

COV	$n = 50$ $\pi = 0.5$	$n = 100$ $\pi = 0.5$	$n = 50$ $\pi = 0.75$	$n = 100$ $\pi = 0.75$	$n = 50$ $\pi = 0.5$	$n = 100$ $\pi = 0.5$	$n = 50$ $\pi = 0.75$	$n = 100$ $\pi = 0.75$
	ZHANG 0				ZHANG 1			
X_1	76.5	96.3	46.1	77.9	65.0	92.2	84.0	97.9
X_2	5.1	5.8	5.5	4.6	6.8	6.0	6.0	6.2
X_3	6.3	5.4	5.3	5.7	4.9	6.6	5.5	5.9
X_4	4.4	5.4	4.4	5.2	5.1	4.4	5.7	5.1
X_5	6.4	8.1	5.6	6.1	7.4	8.5	7.2	9.8
X_6	2.8	4.8	1.8	3.0	3.3	4.4	3.8	5.5
X_7	5.2	6.1	4.9	4.8	4.4	4.9	5.1	4.9
X_8	4.0	4.7	3.9	5.1	4.0	5.8	4.3	3.8
X_9	4.7	6.6	5.1	5.9	5.1	5.3	6.1	7.3
X_{10}	4.8	5.3	3.6	4.0	5.4	4.5	4.5	5.5
X_{11}	9.6	12.6	6.5	9.7	7.5	10.9	9.6	14.0
X_{12}	5.3	6.2	4.2	4.0	4.2	5.1	6.6	7.7

NOTE: COV, covariate; ZHANG 0, working model for the hypothesis test described in Zhang et al. (2008) for the $A = 0$ group; ZHANG 1, working model for the hypothesis test described in Zhang et al. (2008) for the $A = 1$ group.

CHAPTER 7

A Doubly Robust Extension of the Mann-Whitney Test to Adjust for Confounding in Observational Studies

7.1 Introduction

In Chapter 6, we noted that the Mann-Whitney U test (Mann and Whitney 1947) (or equivalently the Wilcoxon rank-sum test (Wilcoxon 1945)) is frequently used to evaluate the effect of a novel treatment as compared to placebo or standard treatment. This is especially so when outcome distributions are heavily skewed. Furthermore, we have seen that the Mann-Whitney U test also comes with a useful effect size measure, namely the **marginal probabilistic index** (MPI, see (6.2)) (Acion et al. 2006). With interest in a causal effect, the Mann-Whitney U test is unfortunately confined to the analysis of randomized experiments. In the analysis of observational studies, interpretation of its results is complicated by the presence of confounding and an association between treatment and outcome detected by the standard Mann-Whitney U test (which is evaluated on the scale of the MPI), is not necessarily attributable to a causal effect of the treatment on the outcome of

interest.

In many observational studies however, one typically collects a rich set of covariates, based on expert-knowledge, to adjust for confounding. It is then hoped for that this set contains most relevant confounders for the treatment-outcome associations, referred to as the **no-unmeasured confounders assumption**. It is common practice to evaluate the treatment effect on the scale of a risk difference (or another function of the the mean responses, e.g., risk ratios or odds ratios); see for example Hahn (1998); Korn and Baumrind (1998); Hirano et al. (2003); Bang and Robins (2005). These effects can for instance be obtained by means of a **marginal structural model** (MSM), see Robins (1998); Robins et al. (2000). This problem is also briefly studied in Section 4.5.1 and Section 4.6 in the context of the bias-reduced doubly robust estimation procedure. These methods successfully deal with measured confounding but may become less interpretable in the presence of highly skewed outcome data and moreover do not appropriately deal with severe outliers, potentially distorting the results. In this case, expressing the causal effect on the scale of the MPI may hence be more convenient (Acion et al. 2006). Adjustment for confounding can be implemented by means of regression adjustment via a probabilistic index model (PIM, Thas et al. (2012)), illustrated in Section 6.3.2, from which an estimator of the MPI can be easily obtained via standardization of the PIM predictions (see Section 6.4). Nevertheless, consistency of this procedure solely relies on correct specification of the PIM. PIMs can however be difficult to specify (see Thas et al. (2012)), especially with complicated error distributions. The resulting estimates may hence be quite vulnerable to model misspecification bias.

In this chapter, we will demonstrate how we can implement adjustment for confounding in the context of the MPI by extending the ideas put forward in Chapter 6, where we extended the classical Mann-Whitney U test to an augmented Mann-Whitney U test, that enables covariate adjustment in randomized experiments. Specifically, we will propose a **doubly robust estimator of the MPI**, which will turn out to be consistent if we either correctly specify a working model for the **propensity score**, the conditional distribution of the treatment given the confounders, or a working model for the **conditional probabilistic index** (CPI) of the outcome conditional on treatment and confounders, which was introduced in Section 6.3.2. Moreover, from a small modification of the semiparametric theory

results developed in Section 6.5, it will follow that the proposed doubly robust estimator is **locally efficient**: it has smallest asymptotic variance within the class of all estimators that are consistent and asymptotically normal under a model that assumes a correctly specified propensity score model, provided that the working model for the CPI is also correctly specified; the semiparametric efficiency bound is thus attained locally.

Wu et al. (2013) and Chen et al. (2013) also propose extensions of the classical Mann-Whitney U test that enable adjustment for confounding in observational studies. Wu et al. (2013) propose an IPTW-type estimator for the MPI which is obtained by fitting a **functional response model** (FRM) (Yu et al. 2011), including a logistic regression model for the propensity score. The extension proposed in Chen et al. (2013) is based on nonparametric kernel estimation of the conditional distribution of the outcome given confounders within treatment groups, which then produces estimators of the marginal treatment-specific outcome distributions. This then leads to a Mann-Whitney-type statistic which serves as an estimator for the MPI. However, both methodologies lack general semiparametric efficiency results and the latter cannot deal with high-dimensional covariates. Chen et al. (2013) do suggest one way to overcome this issue by positing a parametric model for the propensity score and using a nonparametric kernel estimator for the conditional distribution of the outcome within treatment groups, given the one-dimensional propensity score estimator (instead of the original covariate vector) in the construction of their adjusted Mann-Whitney-type statistic. However, by summarizing the covariate information in the one-dimensional propensity score, this may result in a loss of information.

In Section 7.2, we will define the MPI in the context of observational studies in terms of the **counterfactual outcomes** and show in Section 7.3 that the no-unmeasured confounders assumption is sufficient for identification of the MPI from the observed data. Based on these identification results, we propose two simple estimators, each relying on a single working model. Because model misspecification bias is a prevailing concern, a doubly robust estimator of the MPI is constructed in Section 7.4, which has the attractive property of being consistent under correct specification of either of these working models. In Section 7.5, we show that the doubly robust estimator is asymptotically normal and derive its asymptotic

variance. We moreover show how to obtain doubly robust standard errors from these results. In Section 7.6, we demonstrate that the doubly robust estimator is locally efficient under a semiparametric model assuming correct specification of the propensity score working model. We end with a discussion in Section 7.7 where we propose different alternative nuisance parameter estimation strategies, constructed to enhance the performance of the doubly robust estimator.

7.2 The MPI in Observational Studies

Consider an observational study where the interest lies in assessing the causal effect of a dichotomous treatment A, for instance $A = 1$ indicating treated and $A = 0$ indicating untreated, on an outcome measure Y. In this chapter, we will mainly focus on a continuous outcome, but the results also remain valid for discrete or dichotomous outcomes. It is common practice to express the causal effect of A on Y as a mean difference of the counterfactual outcomes:

$$E\{Y(1)\} - E\{Y(0)\},$$

which we studied in Section 4.5.1 and Section 4.6, with $Y(a)$ the **counterfactual outcome** for treatment level a, linked to the observed data through the consistency assumption $Y(a) = Y$ iff $A = a$. However, this measure may become less useful in the presence of heavily skewed outcome data. In this case, a more convenient choice may be to express the causal effect of the treatment on the outcome as the **marginal probabilistic index** (MPI) (Acion et al. 2006):

$$\begin{aligned} \nu_0 &= P\{Y(0) \preceq Y^*(1)\} \\ &= P\{Y(0) < Y^*(1)\} + 0.5P\{Y(0) = Y^*(1)\}, \end{aligned} \tag{7.1}$$

which encodes the probability that if one randomly selects two subjects and randomly chooses to treat one (in which case we get to see $Y^*(1)$) and not to treat the other one (in which case we get to see $Y(0)$), the outcome for the untreated subject is lower than the outcome for the treated subject. Adding $0.5P\{Y(0) = Y^*(1)\}$ to $P\{Y(0) < Y^*(0)\}$ in the definition of the MPI appropriately accounts for ties.

For a continuous outcome, $P\{Y(0) = Y^*(1)\} = 0$, in which case the MPI equals $P\{Y(0) < Y^*(1)\}$. In the absence of confounding, (7.1) equals (6.2).

For a given random sample of size n, where $N_1 = \sum_{i=1}^n A_i$ subjects are treated and the remaining $N_0 = n - N_1$ are not treated, a naive estimator for the MPI ν_0 would be the nonparametric estimator U/N_0N_1 based on the **Mann-Whitney test statistic** $U = \sum_{i=1}^n \sum_{j\neq i}(1-A_i)A_jI(Y_i \preceq Y_j)$, with $I(Y_i \preceq Y_j) = I(Y_i < Y_j) + 0.5I(Y_i = Y_j)$ and $I(\cdot)$ the ordinary indicator function. In the presence of confounding, this estimator is however prone to confounding bias and will not consistently estimate the MPI ν_0, and therefore this confounding bias prohibits the use of the classical Mann-Whitney U test to infer a causal effect.

7.3 Identification and Nuisance Working Models

In many observational studies, for every individual $i = 1,\ldots,n$, besides treatment and outcome data (Y_i, A_i), one often also collects a rich set of covariates $\boldsymbol{X}_i$ based on expert-knowledge. The observed data is then given by the i.i.d. sample $\mathcal{O} = \{\boldsymbol{O}_1,\ldots,\boldsymbol{O}_n\}$ with $\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i)$. It is assumed that this set is sufficient to control for confounding in the sense that $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$ for $a \in \{0,1\}$ (no-unmeasured confounders assumption). This untestable assumption is sufficient to identify the parameter of interest ν_0 in terms of the observed data $\boldsymbol{O}$, which we demonstrate in the following two sections.

7.3.1 Regression imputation estimator

Proposition 7.1. *Under the assumption that $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$ for $a \in \{0,1\}$, the MPI (7.1) can be obtained via standardization of the CPI (see Section 6.3.2):*

$$P\{Y(0) \preceq Y^*(1)\} = E\{P(Y \preceq Y^*|A = 0, A^* = 1, \boldsymbol{X}, \boldsymbol{X}^*)\}, \qquad (7.2)$$

with $(Y, A, \boldsymbol{X})$ and $(Y^*, A^*, \boldsymbol{X}^*)$ data of two independent individuals.

Proof. This easily follows from the law of iterated expectation,

$$\begin{aligned} P\{Y(0) \preceq Y^*(1)\} &= E\big(E[I\{Y(0) \preceq Y^*(1)\}|\boldsymbol{X},\boldsymbol{X}^*]\big) \\ &= E\big(E[I\{Y(0) \preceq Y^*(1)\}|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*]\big) \\ &= E[E\{I(Y \preceq Y^*)|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*\}] \\ &= E\{P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*)\}, \end{aligned}$$

where we use independence of both individuals and where specifically, the second equality follows from the assumption that $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$, and the third equality follows from the consistency assumption. □

Proposition 7.1 suggests one way to obtain an estimator of ν_0, based on standardizing CPI-values. Nonparametric estimation of the CPI

$$m_0(A,A^*,\boldsymbol{X},\boldsymbol{X}^*) = P(Y \preceq Y^*|A,A^*,\boldsymbol{X},\boldsymbol{X}^*)$$

however is infeasible in realistic settings when the covariates contain multiple continuous components, because of the curse of dimensionality (Robins and Ritov 1997). To obtain a well-behaved estimator of the target parameter ν_0, we will therefore postulate a parametric working model for the CPI by means of a **probabilistic index model** (PIM), introduced by Thas et al. (2012) and briefly reviewed in Section 6.3.2. For this purpose, let $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$ denote a working model for the CPI $m_0(A,A^*,\boldsymbol{X},\boldsymbol{X}^*)$, indexed by an r-dimensional parameter $\boldsymbol{\tau}$ where m is a known function, smooth in $\boldsymbol{\tau}$. E.g., we can use model (6.5) or model (6.6) from Section 6.3.2. Let $\hat{\boldsymbol{\tau}}_n$ be an estimator of $\boldsymbol{\tau}$, for instance obtained via the strategy outlined in Appendix 6.A, with probability limit $\boldsymbol{\tau}^*$. Next, we let $\mathcal{M}(\boldsymbol{\tau})$ denote the statistical model for the observed data distribution implied by the working model $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$. Under this model $\mathcal{M}(\boldsymbol{\tau})$, we let $\boldsymbol{\tau}_0$ be such that $m_0(A,A^*,\boldsymbol{X},\boldsymbol{X}^*) = m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}_0)$, in which case we demand that $\boldsymbol{\tau}^* = \boldsymbol{\tau}_0$. A **regression imputation-type estimator** for the MPI can now be obtained by standardizing PIM-predictions:

$$\hat{\nu}_{n,\text{IMP}} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n), \tag{7.3}$$

The estimator $\hat{\nu}_{n,\text{IMP}}$ consistently estimates ν_0 under correct specification of model $\mathcal{M}(\boldsymbol{\tau})$.

7.3.2 IPTW estimator

Instead of exploiting outcome-covariate associations, we can alternatively exploit the treatment-covariate associations.

Proposition 7.2. *Under the assumption that $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$ for $a \in \{0,1\}$, the MPI (7.1) can be obtained via inverse probability of treatment weighting (IPTW):*

$$P\{Y(0) \preceq Y^*(1)\} = E\left\{\frac{(1-A)A^*I(Y \preceq Y^*)}{P(A=0|\boldsymbol{X})P(A^*=1|\boldsymbol{X}^*)}\right\}, \tag{7.4}$$

with $(Y,A,\boldsymbol{X})$ and $(Y^*,A^*,\boldsymbol{X}^*)$ data of two independent individuals.

Proof. For two randomly chosen individuals, we have that

$$\begin{aligned} 1 &= \frac{E(1-A|\boldsymbol{X})E(A^*|\boldsymbol{X}^*)}{P(A=0|\boldsymbol{X})P(A^*=1|\boldsymbol{X}^*)} \\ &= E\left\{\frac{1-A}{P(A=0|\boldsymbol{X})}\frac{A^*}{P(A^*=1|\boldsymbol{X}^*)}\middle|\boldsymbol{X},\boldsymbol{X}^*\right\}. \end{aligned}$$

This implies that

$$\begin{aligned} &P\{Y(0) \preceq Y^*(1)\} \\ &\quad= E\big(E[I\{Y(0) \preceq Y^*(1)\}|\boldsymbol{X},\boldsymbol{X}^*]\big) \\ &\quad= E\left(E\left\{\frac{1-A}{P(A=0|\boldsymbol{X})}\frac{A^*}{P(A^*=1|\boldsymbol{X}^*)}\middle|\boldsymbol{X},\boldsymbol{X}^*\right\}E[I\{Y(0) \preceq Y^*(1)\}|\boldsymbol{X},\boldsymbol{X}^*]\right) \\ &\quad= E\left(E\left[\frac{(1-A)A^*I\{Y(0) \preceq Y^*(1)\}}{P(A=0|\boldsymbol{X})P(A^*=1|\boldsymbol{X}^*)}\middle|\boldsymbol{X},\boldsymbol{X}^*\right]\right) \\ &\quad= E\left[\frac{(1-A)A^*I\{Y(0) \preceq Y^*(1)\}}{P(A=0|\boldsymbol{X})P(A^*=1|\boldsymbol{X}^*)}\right] \\ &\quad= E\left\{\frac{(1-A)A^*I(Y \preceq Y^*)}{P(A=0|\boldsymbol{X})P(A^*=1|\boldsymbol{X}^*)}\right\}, \end{aligned}$$

where the third equality follows from the assumption that $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$, and the last equality follows from the consistency assumption. □

Proposition 7.4 suggests how an IPTW estimator for ν_0 can be obtained, based on the propensity score

$$\pi_0(\boldsymbol{X}) = P(A = 1|\boldsymbol{X}),$$

for which we assume positivity ($1 > 1-\delta \geq \pi_0(\boldsymbol{X}) \geq \delta > 0$ with probability one, see van der Laan and Rose (2011), chap. 10). Similar as for the CPI, we need to posit a parametric working model for the propensity score to obtain a well-behaved estimator of the target parameter ν_0. Let $\pi(\boldsymbol{X};\boldsymbol{\psi})$ denote a working model for the propensity score $\pi_0(\boldsymbol{X})$, indexed by an s-dimensional parameter $\boldsymbol{\psi}$, where π is a known function, smooth in $\boldsymbol{\psi}$, e.g., a logistic regression model $\pi(\boldsymbol{X};\boldsymbol{\psi}) = \text{expit}\{\boldsymbol{\psi}^T\boldsymbol{l}(\boldsymbol{X})\}$ with $\boldsymbol{l} = (1, l_1, \ldots, l_{s-1})$. Let $\hat{\boldsymbol{\psi}}_n$ be an estimator of $\boldsymbol{\psi}$, for instance the MLE, and let $\boldsymbol{\psi}^*$ denote the corresponding probability limit. Let $\mathcal{M}(\boldsymbol{\psi})$ denote the statistical model for the observed data distribution implied by the working model $\pi(\boldsymbol{X};\boldsymbol{\psi})$. Under this model $\mathcal{M}(\boldsymbol{\psi})$, we let $\boldsymbol{\psi}_0$ be such that $\pi_0(\boldsymbol{X}) = \pi(\boldsymbol{X};\boldsymbol{\psi}_0)$, in which case we demand that $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$. An **IPTW estimator** for the MPI can then be obtained as

$$\hat{\nu}_{n,\text{IPTW}} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{(1-A_i)A_j I(Y_i \preceq Y_j)}{\{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)\}\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n)}. \tag{7.5}$$

Under model $\mathcal{M}(\boldsymbol{\psi})$, the estimator $\hat{\nu}_{n,\text{IPTW}}$ will consistently estimate ν_0. A modification of this IPTW estimator is also suggested in the discussion by Stijn Vansteelandt (Vansteelandt 2012) of the read paper Thas et al. (2012), see equation (40), so as to make the IPTW estimator **sample bounded** (SB) (Robins et al. 2007, sec. 4.1), which is accomplished by dividing by the sample mean of the inverse weights:

$$\hat{\nu}^{(\text{SB})}_{n,\text{IPTW}} = \frac{\sum_{i=1}^{n}\sum_{j\neq i}(1-A_i)A_j I(Y_i \preceq Y_j)/[\{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)\}\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n)]}{\sum_{i=1}^{n}\sum_{j\neq i}(1-A_i)A_j/[\{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n)\}\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n)]}.$$

In this discussion, it is also argued that in practice, the propensity score model might be easier to specify than the dependence of the probabilistic index on the covariates $\boldsymbol{X}$. This is because the associations of the covariates with the treatment

are often better understood than the associations of the covariates with the outcome. Moreover, complicated error distributions for the outcome typically imply complex PIMs (see Thas et al. (2012)), making them more likely to be misspecified. This IPTW estimator is also suggested in Chen et al. (2013) in the context of a randomized experiment where the outcome Y is susceptible to missingness explainable by the measured covariates $\boldsymbol{X}$.

7.4 Doubly Robust Estimation of the MPI

We argued in Section 3.3 that a prevailing concern is that misspecification of these nuisance working models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\tau})$ may induce bias in the estimator of the target parameter ν_0. We also noted that this concern of model misspecification bias can often be lessened via the use of doubly robust estimators. In this section, we will therefore propose such a doubly robust estimator for the MPI ν_0, which will consistently estimate the target parameter ν_0 under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\tau})$; that is, when we either correctly specify the working model for the propensity score or the working model for the CPI.

Based on the working models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\tau})$ and the notation introduced in the previous section and with the **pseudo-observations** $I_{ij} = I(Y_i \preceq Y_j)$ for individuals i and j, define the estimator

$$
\begin{aligned}
&\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) \qquad (7.6)\\
&= \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \left[\frac{1-A_i}{1-\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)} \frac{A_j}{\pi(\boldsymbol{X}_j; \hat{\boldsymbol{\psi}}_n)} I_{ij} \right.\\
&\qquad \left. + \left\{ 1 - \frac{1-A_i}{1-\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)} \frac{A_j}{\pi(\boldsymbol{X}_j; \hat{\boldsymbol{\psi}}_n)} \right\} m(0,1,\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\boldsymbol{\tau}}_n) \right]\\
&= \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \left[m(0,1,\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\boldsymbol{\tau}}_n) \right.\\
&\qquad \left. + \frac{1-A_i}{1-\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\psi}}_n)} \frac{A_j}{\pi(\boldsymbol{X}_j; \hat{\boldsymbol{\psi}}_n)} \{ I_{ij} - m(0,1,\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\boldsymbol{\tau}}_n) \} \right].
\end{aligned}
$$

From the above expressions of $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$, it is seen that this estimator forms a

compromise between the IPTW estimator $\hat{\nu}_{n,\text{IPTW}}$ (that solely relies on $\mathcal{M}(\boldsymbol{\psi})$) and the regression imputation-based estimator $\hat{\nu}_{n,\text{IMP}}$ (that solely relies on $\mathcal{M}(\boldsymbol{\tau})$). It follows however from the following theorem that $\hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ will be consistent under correct specification of either of these working models.

Theorem 7.1. *The estimator $\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ is a consistent estimator of ν_0 under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\tau})$; that is, it is **doubly robust**.*

Proof. With $\boldsymbol{\tau}^*$ the probability limit of $\hat{\boldsymbol{\tau}}_n$ and $\boldsymbol{\psi}^*$ the probability limit of $\hat{\boldsymbol{\psi}}_n$, double robustness of the estimator $\hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ follows if we can show that

$$E\left[\frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}^*)}\frac{A^*}{\pi(\boldsymbol{X}^*;\boldsymbol{\psi}^*)}I(Y \preceq Y^*) + \left\{1 - \frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}^*)}\frac{A^*}{\pi(\boldsymbol{X}^*;\boldsymbol{\psi}^*)}\right\} m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\right] = \nu_0$$

when either $\mathcal{M}(\boldsymbol{\psi})$ or $\mathcal{M}(\boldsymbol{\tau})$ holds.

(a) **Suppose $\mathcal{M}(\boldsymbol{\psi})$ is correctly specified.** In this case, $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$ and thus $\pi(\boldsymbol{X};\boldsymbol{\psi}^*) = \pi_0(\boldsymbol{X})$. From Section 7.3.2, we already know that

$$E\left\{\frac{1-A}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}I(Y \preceq Y^*)\right\} = \nu_0.$$

Next, from the law of iterated expectation, it follows that

$$\begin{aligned}
&E\left[\left\{1 - \frac{1-A}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}\right\} m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\right] \\
&\quad = E\left(E\left[\left\{1 - \frac{1-A}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}\right\} m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\middle|\boldsymbol{X},\boldsymbol{X}^*\right]\right) \\
&\quad = E\left[\left\{1 - \frac{E(1-A|\boldsymbol{X})}{1-\pi_0(\boldsymbol{X})}\frac{E(A^*|\boldsymbol{X}^*)}{\pi_0(\boldsymbol{X}^*)}\right\} m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\right] \\
&\quad = 0.
\end{aligned}$$

(b) **Suppose $\mathcal{M}(\boldsymbol{\tau})$ is correctly specified.** In this case, $\boldsymbol{\tau}^* = \boldsymbol{\tau}_0$ and thus

$m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*) = m_0(0,1,\boldsymbol{X},\boldsymbol{X}^*)$. From Section 7.3.1, we know that

$$E\{m_0(0,1,\boldsymbol{X},\boldsymbol{X}^*)\} = \nu_0.$$

Next, from the law of iterated expectation, it follows that

$$\begin{aligned}
&E\left[\frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}^*)}\frac{A^*}{\pi(\boldsymbol{X}^*;\boldsymbol{\psi}^*)}\{I(Y \preceq Y^*) - m_0(0,1,\boldsymbol{X},\boldsymbol{X}^*)\}\right] \\
&= E\left(E\left[\frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}^*)}\frac{A^*}{\pi(\boldsymbol{X}^*;\boldsymbol{\psi}^*)}\right.\right. \\
&\qquad \left.\left.\times\{I(Y \preceq Y^*) - m_0(0,1,\boldsymbol{X},\boldsymbol{X}^*)\}\,\middle|\, A,A^*,\boldsymbol{X},\boldsymbol{X}^*\right]\right) \\
&= E\left(\frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}^*)}\frac{A^*}{\pi(\boldsymbol{X}^*;\boldsymbol{\psi}^*)}\right. \\
&\qquad \left.\times E\left[E\left\{I(Y \preceq Y^*)|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*\right\} - m_0(0,1,\boldsymbol{X},\boldsymbol{X}^*)\right]\right) \\
&= 0.
\end{aligned}$$

We can conclude that $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ consistently estimates ν_0 under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\tau})$. □

Remark 7.1. *When treatment is randomized and the propensity equals a constant π and it is estimated via MLE (that is, $\hat{\pi}_n = n^{-1}\sum_{i=1}^n A_i$), the doubly robust estimator $\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ (7.6) reduces to the estimator $\hat{\nu}_{n,adap}$, see equation (6.17), developed in Chapter 6 to increase the efficiency of the unadjusted estimator for the MPI in randomized experiments.*

7.5 Asymptotic Distribution of The Doubly Robust Estimator

In this section, we will establish, under suitable regularity conditions, the asymptotic linearity of the doubly robust estimator $\hat{v}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ by deriving its influence function, from which asymptotic normality and an expression for the asymptotic variance of the estimator will follow. We also provide an asymptotic variance estimator, enabling doubly robust inference.

Estimating functions for the nuisance working models

For the working model $\pi(\boldsymbol{X};\boldsymbol{\psi})$, inducing the statistical model $\mathcal{M}(\boldsymbol{\psi})$, we let $\hat{\boldsymbol{\psi}}_n$ be an estimator, with probability limit $\boldsymbol{\psi}^*$, defined as the solution to the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\hat{\boldsymbol{\psi}}_n) = \mathbf{0},$$

where the estimating function $\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi})$ is of the form

$$\boldsymbol{d}_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi})\{A-\pi(\boldsymbol{X};\boldsymbol{\psi})\},$$

with $\boldsymbol{d}_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi})$ an arbitrary index function (depending only on the covariates $\boldsymbol{X}$ and potentially also $\boldsymbol{\psi}$) and the estimating function thus satisfies $E\{\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)\} = \mathbf{0}$, where $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$ under $\mathcal{M}(\boldsymbol{\psi})$. For instance, for the MLE, $\boldsymbol{d}_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}) = [\{1-\pi(\boldsymbol{X};\boldsymbol{\psi})\}\pi(\boldsymbol{X};\boldsymbol{\psi})]^{-1}\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi})$ with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}) = \partial\pi(\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}$.

For the CPI-working model $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$, inducing the model $\mathcal{M}(\boldsymbol{\tau})$, let $\hat{\boldsymbol{\tau}}_n$ be an estimator, with probability limit $\boldsymbol{\tau}^*$, defined as the solution to the estimating equation

$$\sum_{i=1}^{n}\sum_{i\neq j} \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i,\boldsymbol{O}_j;\hat{\boldsymbol{\tau}}_n) = \mathbf{0},$$

where the estimating function $\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau})$ is of the form

$$\boldsymbol{d}_{\boldsymbol{\tau}}(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})\{I(Y\preceq Y^*) - m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})\},$$

with $\boldsymbol{d}_{\boldsymbol{\tau}}(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$ an arbitrary index function (depending only on the treat-

ment and covariates and potentially also $\boldsymbol{\tau}$) and the estimating function thus satisfies $E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)\}=\mathbf{0}$ where $\boldsymbol{\tau}^*=\boldsymbol{\tau}_0$ under $\mathcal{M}(\boldsymbol{\tau})$. For instance, when using the estimating equation (6.22),

$$\boldsymbol{d}_{\boldsymbol{\tau}}(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})=\frac{m_{\boldsymbol{\tau}}(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})}{\boldsymbol{V}\{m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})\}}$$

with gradient $m_{\boldsymbol{\tau}}(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})=\partial m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})/\partial\boldsymbol{\tau}$ and conditional (model-based) variance $\boldsymbol{V}\{m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})\}=\mathrm{var}\{I(Y\preceq Y^*)|A,A^*,\boldsymbol{X},\boldsymbol{X}^*\}$.

Stochastic equicontinuity condition

For given working models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\tau})$, define the function

$$\begin{aligned}U(\boldsymbol{O}_i,\boldsymbol{O}_j;\nu,\boldsymbol{\psi},\boldsymbol{\tau}) &= m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})-\nu\\ &\quad+\frac{1-A_i}{1-\pi(\boldsymbol{X}_i;\boldsymbol{\psi})}\frac{A_j}{\pi(\boldsymbol{X}_j;\boldsymbol{\psi})}\{I_{ij}-m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})\}.\end{aligned}$$

Next, define the U-statistic

$$U_n(\nu,\boldsymbol{\psi},\boldsymbol{\tau})=\{n(n-1)\}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\{U(\boldsymbol{O}_i,\boldsymbol{O}_j;\nu,\boldsymbol{\psi},\boldsymbol{\tau})+U(\boldsymbol{O}_j,\boldsymbol{O}_i;\nu,\boldsymbol{\psi},\boldsymbol{\tau})\}/2,$$

where we symmetrize the function U so to make it permutation symmetric in its arguments $\boldsymbol{O}_i$ and $\boldsymbol{O}_j$, simplifying the formulas in the derivation of its asymptotic distribution (see for instance Chapter 12 of van der Vaart (1998)). It follows that the doubly robust estimator $\hat{\nu}_{n,\mathrm{DR}}\equiv\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ is the solution to

$$U_n(\hat{\nu}_{n,\mathrm{DR}},\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)=0. \tag{7.7}$$

With $\mathcal{U}(\nu,\boldsymbol{\psi},\boldsymbol{\tau})=E[\{U(\boldsymbol{O}_i,\boldsymbol{O}_j;\nu,\boldsymbol{\psi},\boldsymbol{\tau})+U(\boldsymbol{O}_j,\boldsymbol{O}_i;\nu,\boldsymbol{\psi},\boldsymbol{\tau})\}/2]$, we will assume that the U-process (that is, the empirical process associated with the U-statistic $U_n(\nu,\boldsymbol{\psi},\boldsymbol{\tau})$) $\{\mathbb{U}_n(\nu,\boldsymbol{\psi},\boldsymbol{\tau}):n\geq 1,\boldsymbol{\theta}=(\nu,\boldsymbol{\psi}^T,\boldsymbol{\tau}^T)^T\in\Theta\}$ for a compact set $\Theta\subset\mathbb{R}^{1+s+r}$, with

$$\mathbb{U}_n(\nu,\boldsymbol{\psi},\boldsymbol{\tau})=n^{1/2}\{U_n(\nu,\boldsymbol{\psi},\boldsymbol{\tau})-\mathcal{U}(\nu,\boldsymbol{\psi},\boldsymbol{\tau})\} \tag{7.8}$$

is **stochastically equicontinuous**, which we define below.

Definition 7.1 (Stochastically Equicontinuous)**.** *The U-process $\{\mathbb{U}_n(\boldsymbol{\theta}) : n \geq 1, \boldsymbol{\theta} \in \Theta\}$ with Θ a compact set is called **stochastically equicontinuous** if for every $\varepsilon > 0$ and $\eta > 0$, there exists a $\delta > 0$ such that*

$$\limsup_{n\to\infty} P\left[\sup_{\substack{\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|<\delta \\ \boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Theta}} \left|\mathbb{U}_n(\boldsymbol{\theta}_1) - \mathbb{U}_n(\boldsymbol{\theta}_2)\right| > \eta\right] < \varepsilon, \tag{7.9}$$

with $\|\cdot\|$ and appropriate norm, e.g., the Euclidean norm for the $(1+s+r)$-dimensional parameter $\boldsymbol{\theta} = (\nu, \boldsymbol{\psi}^T, \boldsymbol{\tau}^T)^T$.

Stochastic equicontinuity of the U-process $\{\mathbb{U}_n(\boldsymbol{\theta}) : n \geq 1, \boldsymbol{\theta} \in \Theta\}$ states that the function $\mathbb{U}_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ uniformly over Θ with high probability and for n large. Sufficient conditions for stochastic equicontinuity can be derived from Nolan and Pollard (1987, 1988) and for a detailed discussion and examples, we refer to Andrews (1994a,b).

7

Asymptotic linearity and asymptotic distribution

Throughout, we will assume sufficient regularity conditions, see Appendix B of Robins et al. (1994) and the Appendix of Rotnitzky et al. (2006). By construction of the doubly robust estimator $\hat{\nu}_{n,\mathrm{DR}}$, we have that

$$\begin{aligned}
0 &= n^{1/2} U_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) \\
&= n^{1/2}\Big\{U_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathcal{U}(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)\Big\} \\
&\quad - n^{1/2}\Big\{U_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) - \mathcal{U}(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)\Big\} \\
&\quad + n^{1/2}\Big\{\mathcal{U}(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathcal{U}(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)\Big\} + n^{1/2} U_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) \\
&= \mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) \\
&\quad + n^{1/2}\Big\{\mathcal{U}(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathcal{U}(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)\Big\} + n^{1/2} U_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*).
\end{aligned}$$

Because we assume that the U-process $\mathbb{U}_n(\nu, \boldsymbol{\psi}, \boldsymbol{\tau})$ is stochastically equicontinuous, it follows that $\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) = o_p(1)$. Indeed, define $\hat{\boldsymbol{\theta}}_n =$

$(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n^T, \hat{\boldsymbol{\tau}}_n^T)^T$ and $\boldsymbol{\theta}^* = (\nu_0, \boldsymbol{\psi}^{*,T}, \boldsymbol{\tau}^{*,T})^T$, it then follows from (7.9) that for any $\varepsilon > 0$ and $\eta > 0$, we can find a $\delta > 0$ such that

$$\begin{aligned}
&\limsup_{n\to\infty} P\Big\{ |\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)| > \eta \Big\} \\
&\leq \limsup_{n\to\infty} P\Big[\Big\{ |\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)| > \eta \Big\} \cap \Big\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| < \delta \Big\} \Big] \\
&\quad + \limsup_{n\to\infty} P\Big[\Big\{ |\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)| > \eta \Big\} \cap \Big\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \geq \delta \Big\} \Big] \\
&\leq \limsup_{n\to\infty} P\Big[\Big\{ |\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)| > \eta \Big\} \cap \Big\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| < \delta \Big\} \Big] \\
&\quad + \limsup_{n\to\infty} P\Big\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \geq \delta \Big\}.
\end{aligned}$$

Because $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^*$, it follows that $\limsup_{n\to\infty} P\{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \geq \delta\} = 0$. Hence,

$$\begin{aligned}
&\limsup_{n\to\infty} P\Big\{ |\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)| > \eta \Big\} \\
&\leq \limsup_{n\to\infty} P\Big[\Big\{ |\mathbb{U}_n(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathbb{U}_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)| > \eta \Big\} \cap \Big\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| < \delta \Big\} \Big] \\
&\leq \limsup_{n\to\infty} P\left\{ \sup_{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| < \delta} |\mathbb{U}_n(\nu_1, \boldsymbol{\psi}_1, \boldsymbol{\tau}_1) - \mathbb{U}_n(\nu_2, \boldsymbol{\psi}_2, \boldsymbol{\tau}_2)| > \eta \right\} < \varepsilon,
\end{aligned}$$

where the last inequality follows from the stochastic equicontinuity condition. We thus have that

$$0 = o_p(1) + n^{1/2}\Big\{ \mathcal{U}(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \mathcal{U}(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) \Big\} + n^{1/2} U_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*).$$

Next, consider the Taylor expansion

$$\begin{aligned}
n^{1/2}\mathcal{U}(\hat{\nu}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) = n^{1/2}\mathcal{U}(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) \\
+ \mathcal{U}_\nu(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) n^{1/2}(\hat{\nu}_{n,\mathrm{DR}} - \nu_0) \\
+ \mathcal{U}_{\boldsymbol{\psi}}^T(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) \\
+ \mathcal{U}_{\boldsymbol{\tau}}^T(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*),
\end{aligned}$$

with derivatives (assuming they exist)

$$\begin{aligned}
\mathcal{U}_\nu(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) &= \partial \mathcal{U}(\nu, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n)/\partial \nu|_{\nu=\tilde{\nu}_n}, \\
\mathcal{U}_{\boldsymbol{\psi}}(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) &= \partial \mathcal{U}(\tilde{\nu}_n, \boldsymbol{\psi}, \tilde{\boldsymbol{\tau}}_n)/\partial \boldsymbol{\psi}|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}_n}, \\
\mathcal{U}_{\boldsymbol{\tau}}(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) &= \partial \mathcal{U}(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \boldsymbol{\tau})/\partial \boldsymbol{\tau}|_{\boldsymbol{\tau}=\tilde{\boldsymbol{\tau}}_n},
\end{aligned}$$

and with the values $\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n$ intermediate on the line segment connecting $\hat{\nu}_{n,\mathrm{DR}}$ and ν_0, $\hat{\boldsymbol{\psi}}_n$ and $\boldsymbol{\psi}^*$, and $\hat{\boldsymbol{\tau}}_n$ and $\boldsymbol{\tau}^*$, respectively. Because $\mathcal{U}_\nu(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) = -1$, we find that

$$\begin{aligned}
n^{1/2}(\hat{\nu}_{n,\mathrm{DR}} - \nu_0) = {} & \mathcal{U}_{\boldsymbol{\psi}}^T(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) + \mathcal{U}_{\boldsymbol{\tau}}^T(\tilde{\nu}_n, \tilde{\boldsymbol{\psi}}_n, \tilde{\boldsymbol{\tau}}_n) n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*) \\
& + n^{1/2} U_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) + o_p(1).
\end{aligned}$$

Under sufficient regularity, we then find that

$$\begin{aligned}
n^{1/2}(\hat{\nu}_{n,\mathrm{DR}} - \nu_0) = {} & \mathcal{U}_{\boldsymbol{\psi}}^T(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) + \mathcal{U}_{\boldsymbol{\tau}}^T(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*) \\
& + n^{1/2} U_n(\nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) + o_p(1), \qquad (7.10)
\end{aligned}$$

because $\tilde{\nu}_n \xrightarrow{p} \nu_0$, $\tilde{\boldsymbol{\psi}}_n \xrightarrow{p} \boldsymbol{\psi}^*$ and $\tilde{\boldsymbol{\tau}}_n \xrightarrow{p} \boldsymbol{\tau}^*$ since $\hat{\nu}_{n,\mathrm{DR}}$, $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\tau}}_n$ are consistent estimators for ν_0, $\boldsymbol{\psi}^*$ and $\boldsymbol{\tau}^*$ respectively.

The next step is to invoke the asymptotic linearity of the nuisance parameter estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\tau}}_n$. For $\hat{\boldsymbol{\psi}}_n$, it follows from standard regularity conditions that

$$n^{1/2}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) = n^{-1/2} \sum_{i=1}^{n} \left[-E\left\{ \frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}; \boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^T} \right\}^{-1} \right] \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \boldsymbol{\psi}^*) + o_p(1) \qquad (7.11)$$

and thus that the estimator $\hat{\boldsymbol{\psi}}_n$ is asymptotically linear with influence function $-E\{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}; \boldsymbol{\psi}^*)/\partial \boldsymbol{\psi}^T\}^{-1} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}; \boldsymbol{\psi}^*)$. To demonstrate asymptotic linearity of $\hat{\boldsymbol{\tau}}_n$, we also need to impose a stochastic equicontinuity condition. For this purpose, let

$$\boldsymbol{K}_n(\boldsymbol{\tau}) = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \boldsymbol{\tau}) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j, \boldsymbol{O}_i; \boldsymbol{\tau})\}/2.$$

It follows that $\hat{\boldsymbol{\tau}}_n$ solves $\boldsymbol{K}_n(\hat{\boldsymbol{\tau}}_n) = \boldsymbol{0}$. Next define $\mathcal{K}(\boldsymbol{\tau}) = E[\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \boldsymbol{\tau}) +$

$\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j,\boldsymbol{O}_i;\boldsymbol{\tau})\}/2]$. We now assume that the U-process $\{\mathbb{K}_n(\boldsymbol{\tau}) : n \geq 1, \boldsymbol{\tau} \in \mathbb{T}\}$, for a compact set $\mathbb{T} \subset \mathbb{R}^r$ and with $\mathbb{K}_n(\boldsymbol{\tau}) = n^{1/2}\{\boldsymbol{K}_n(\boldsymbol{\tau}) - \mathcal{K}(\boldsymbol{\tau})\}$, is stochastically equicontinuous. From a similar reasoning as before, it follows that

$$\begin{aligned}
\mathbf{0} &= n^{1/2}\boldsymbol{K}_n(\hat{\boldsymbol{\tau}}_n) \\
&= n^{1/2}\Big\{\boldsymbol{K}_n(\hat{\boldsymbol{\tau}}_n) - \mathcal{K}(\hat{\boldsymbol{\tau}}_n)\Big\} - n^{1/2}\Big\{\boldsymbol{K}_n(\boldsymbol{\tau}^*) - \mathcal{K}(\boldsymbol{\tau}^*)\Big\} \\
&\quad + n^{1/2}\Big\{\mathcal{K}(\hat{\boldsymbol{\tau}}_n) - \mathcal{K}(\boldsymbol{\tau}^*)\Big\} + n^{1/2}\boldsymbol{K}_n(\boldsymbol{\tau}^*) \\
&= \mathbb{K}_n(\hat{\boldsymbol{\tau}}_n) - \mathbb{K}_n(\boldsymbol{\tau}^*) + n^{1/2}\Big\{\mathcal{K}(\hat{\boldsymbol{\tau}}_n) - \mathcal{K}(\boldsymbol{\tau}^*)\Big\} + n^{1/2}\boldsymbol{K}_n(\boldsymbol{\tau}^*) \\
&= o_p(1) + n^{1/2}\Big\{\mathcal{K}(\hat{\boldsymbol{\tau}}_n) - \mathcal{K}(\boldsymbol{\tau}^*)\Big\} + n^{1/2}\boldsymbol{K}_n(\boldsymbol{\tau}^*),
\end{aligned}$$

where the last equality follows from the stochastic equicontinuity. Next, consider the Taylor expansion

$$n^{1/2}\mathcal{K}(\hat{\boldsymbol{\tau}}_n) = n^{1/2}\mathcal{K}(\boldsymbol{\tau}^*) + \mathcal{K}^T_{\boldsymbol{\tau}}(\tilde{\boldsymbol{\tau}}_n)n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*),$$

where $\mathcal{K}_{\boldsymbol{\tau}}(\tilde{\boldsymbol{\tau}}_n) = \partial\mathcal{K}(\boldsymbol{\tau})/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\tilde{\boldsymbol{\tau}}_n}$ and $\tilde{\boldsymbol{\tau}}_n$ is intermediate on the line segment connecting $\hat{\boldsymbol{\tau}}_n$ and $\boldsymbol{\tau}^*$. Because $\tilde{\boldsymbol{\tau}}_n \xrightarrow{p} \boldsymbol{\tau}^*$ (since $\hat{\boldsymbol{\tau}}_n$ is a consistent estimator of $\boldsymbol{\tau}^*$), we find that under sufficient regularity,

$$\begin{aligned}
n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*) &= -\left\{\mathcal{K}^T_{\boldsymbol{\tau}}(\tilde{\boldsymbol{\tau}}_n)\right\}^{-1} n^{1/2}\boldsymbol{K}_n(\boldsymbol{\tau}^*) + o_p(1) \\
&= -\left\{\mathcal{K}^T_{\boldsymbol{\tau}}(\boldsymbol{\tau}^*)\right\}^{-1} n^{1/2}\boldsymbol{K}_n(\boldsymbol{\tau}^*) + o_p(1).
\end{aligned}$$

To establish the asymptotic linearity of $\hat{\boldsymbol{\tau}}_n$, we finally need to consider the Hájek projection (Hájek 1970; van der Vaart 1998) of $\boldsymbol{K}_n(\boldsymbol{\tau}^*)$. This is given by ($i \neq j$)

$$\begin{aligned}
\widehat{\boldsymbol{K}}_n(\boldsymbol{\tau}^*) &= \sum_{i=1}^{n} E\{\boldsymbol{K}_n(\boldsymbol{\tau}^*)|\boldsymbol{O}_i\} \\
&= n^{-1}\sum_{i=1}^{n} E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i,\boldsymbol{O}_j;\boldsymbol{\tau}^*) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j,\boldsymbol{O}_i;\boldsymbol{\tau}^*)|\boldsymbol{O}_i\}.
\end{aligned}$$

From Theorem 12.3 in van der Vaart (1998), it follows that the Hájek projection

satisfies $n^{1/2}\{\boldsymbol{K}_n(\boldsymbol{\tau}^*) - \widehat{\boldsymbol{K}}_n(\boldsymbol{\tau}^*)\} = o_p(1)$, so that

$$
\begin{aligned}
&n^{1/2}(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^*) \\
&= n^{-1/2}\sum_{i=1}^{n}\left[-\left\{\mathcal{K}_{\boldsymbol{\tau}}^T(\boldsymbol{\tau}^*)\right\}^{-1}\right] E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i,\boldsymbol{O}_j;\boldsymbol{\tau}^*) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j,\boldsymbol{O}_i;\boldsymbol{\tau}^*)|\boldsymbol{O}_i\} + o_p(1),
\end{aligned}
\tag{7.12}
$$

from which we can conclude that $\hat{\boldsymbol{\tau}}_n$ is asymptotically linear with influence function $-\left\{\mathcal{K}_{\boldsymbol{\tau}}^T(\boldsymbol{\tau}^*)\right\}^{-1} E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}^*,\boldsymbol{O};\boldsymbol{\tau}^*)|\boldsymbol{O}\}$.

Using (7.11) and (7.12) in (7.10) and using the definitions of the functions $\mathcal{U}(\nu,\boldsymbol{\psi},\boldsymbol{\tau})$ and $\mathcal{K}(\boldsymbol{\tau})$, and assuming sufficient regularity so we can interchange differentiation and integration, we find that (with $i \neq j$)

$$
\begin{aligned}
&n^{1/2}(\hat{\nu}_{n,\mathrm{DR}} - \nu_0) \\
&\quad = n^{1/2}U_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*) + o_p(1) \\
&\qquad -\mathcal{U}_{\boldsymbol{\psi}}^T(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)n^{-1/2}\sum_{i=1}^{n}\left[E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^T}\right\}^{-1}\right]\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\boldsymbol{\psi}^*) \\
&\qquad -\mathcal{U}_{\boldsymbol{\tau}}^T(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)n^{-1/2}\sum_{i=1}^{n}\left[\left\{\mathcal{K}_{\boldsymbol{\tau}}^T(\boldsymbol{\tau}^*)\right\}^{-1}\right] \\
&\qquad\qquad \times E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i,\boldsymbol{O}_j;\boldsymbol{\tau}^*) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j,\boldsymbol{O}_i;\boldsymbol{\tau}^*)|\boldsymbol{O}_i\} \\
&\quad = n^{1/2}U_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*) + o_p(1) \\
&\qquad -E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\psi}^T}\right\}n^{-1/2}\sum_{i=1}^{n}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^T}\right\}^{-1}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i;\boldsymbol{\psi}^*) \\
&\qquad -E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}n^{-1/2}\sum_{i=1}^{n}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}^{-1} \\
&\qquad\qquad \times E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i,\boldsymbol{O}_j;\boldsymbol{\tau}^*) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j,\boldsymbol{O}_i;\boldsymbol{\tau}^*)|\boldsymbol{O}_i\}.
\end{aligned}
$$

A final step is to show the asymptotic linearity of the term $n^{1/2}U_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)$, which is again obtained by considering its Hájek projection ($i \neq j$):

$$
\widehat{U}_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*) \;=\; \sum_{i=1}^{n}E\{U_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)|\boldsymbol{O}_i\}
$$

$$= \quad n^{-1}\sum_{i=1}^{n} E\{U(\boldsymbol{O}_i,\boldsymbol{O}_j;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)+U(\boldsymbol{O}_j,\boldsymbol{O}_i;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)|\boldsymbol{O}_i\}.$$

Because $n^{1/2}\{U_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)-\widehat{U}_n(\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)\}=o_p(1)$ (which follows from Theorem 12.3 of van der Vaart (1998)), it follows that the doubly robust estimator $\hat{\nu}_{n,\mathrm{DR}}$ is asymptotically linear:

$$n^{1/2}(\hat{\nu}_{n,\mathrm{DR}}-\nu_0)=n^{-1/2}\sum_{i=1}^{n}\phi_\nu(\boldsymbol{O}_i;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)+o_p(1), \tag{7.13}$$

with influence function

$$\begin{aligned}
&\phi_\nu(\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*) \qquad (7.14)\\
&= E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)+U(\boldsymbol{O}^*,\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)|\boldsymbol{O}\}\\
&\quad -E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial\boldsymbol{\psi}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}^T}\right\}^{-1}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)\\
&\quad -E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial\boldsymbol{\tau}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)}{\partial\boldsymbol{\tau}^T}\right\}^{-1}\\
&\qquad\qquad \times E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)+\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}^*,\boldsymbol{O};\boldsymbol{\tau}^*)|\boldsymbol{O}\}.
\end{aligned}$$

We may thus conclude with the following theorem:

Theorem 7.2 (Asymptotic Linearity of $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$)**.** *Under suitable regularity conditions (see Appendix B of Robins et al. (1994) and the Appendix of Rotnitzky et al. (2006)), assuming positivity ($1>1-\delta\geq\pi_0(\boldsymbol{X})\geq\delta>0$ with probability one, see van der Laan and Rose (2011), chap. 10) and assuming that the U-processes $\{\mathbb{U}_n(\nu,\boldsymbol{\psi},\boldsymbol{\tau}):n\geq 1,\boldsymbol{\theta}=(\nu,\boldsymbol{\psi}^T,\boldsymbol{\tau}^T)^T\in\Theta\}$ and $\{\mathbb{K}_n(\boldsymbol{\tau}):n\geq 1,\boldsymbol{\tau}\in\mathbb{T}\}$ are stochastically equicontinuous, it follows that under the union model $\mathcal{M}(\boldsymbol{\psi})\cup\mathcal{M}(\boldsymbol{\tau})$, the doubly robust estimator $\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ is asymptotically linear*

$$n^{1/2}\{\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)-\nu_0\}=n^{-1/2}\sum_{i=1}^{n}\phi_\nu(\boldsymbol{O}_i;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)+o_p(1)$$

with influence function

$$\begin{aligned}
&\phi_\nu(\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)\\
&= E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)+U(\boldsymbol{O}^*,\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)|\boldsymbol{O}\}\\
&\quad -E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\psi}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^T}\right\}^{-1}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)\\
&\quad -E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}^{-1}\\
&\qquad\qquad \times E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)+\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}^*,\boldsymbol{O};\boldsymbol{\tau}^*)|\boldsymbol{O}\},
\end{aligned}$$

and $(\boldsymbol{\psi}^{*,T},\boldsymbol{\tau}^{*,T})^T$ *the probability limit of* $(\hat{\boldsymbol{\psi}}_n^T,\hat{\boldsymbol{\tau}}_n^T)^T$*. It follows that*

$$n^{1/2}\{\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)-\nu_0\}\xrightarrow{d}N[0,var\{\phi_\nu(\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)\}],$$

so that the doubly robust estimator is asymptotically normal with asymptotic variance equal to the variance of the influence function $\phi_\nu(\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)$*.*

7

From the expression (7.14) for the influence function of the doubly robust estimator, it also follows that the influence function would equal $E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)+U(\boldsymbol{O}^*,\boldsymbol{O};\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)|\boldsymbol{O}\}$ when the values $\boldsymbol{\psi}^*$ and $\boldsymbol{\tau}^*$ would be known to us. The term

$$-E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\psi}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^T}\right\}^{-1}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}^*)$$

corrects for the estimation of $\boldsymbol{\psi}$ under misspecification of $\mathcal{M}(\boldsymbol{\tau})$ and likewise, the term

$$\begin{aligned}
&-E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}^*,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}^{-1}\\
&\qquad\qquad \times E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O},\boldsymbol{O}^*;\boldsymbol{\tau}^*)+\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}^*,\boldsymbol{O};\boldsymbol{\tau}^*)|\boldsymbol{O}\}
\end{aligned}$$

corrects for the estimation of $\boldsymbol{\tau}$ under misspecification of model $\mathcal{M}(\boldsymbol{\psi})$ (see also Section 3.4 for a discussion on this).

Estimating the asymptotic variance

A consistent estimator of the asymptotic variance of the doubly robust estimator $\hat{v}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ can be obtained via the empirical variance of the estimated influence function $\hat{\phi}_{n,v}\{\boldsymbol{O}; \hat{v}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n), \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$, resulting in the sandwich estimator:

$$\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \hat{\phi}_{n,v}^2\{\boldsymbol{O}_i; \hat{v}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n), \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n\}, \tag{7.15}$$

where (with $\hat{v}_{n,\mathrm{DR}} \equiv \hat{v}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$)

$$\begin{aligned}
&\hat{\phi}_{n,v}(\boldsymbol{O}_i; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) \\
&\quad = \hat{E}_n\{U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) + U(\boldsymbol{O}_j, \boldsymbol{O}_i; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) | \boldsymbol{O}_i\} \\
&\qquad - \hat{E}_n\left\{\frac{\partial U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\psi}^T}\right\} \hat{E}_n\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n)}{\partial \boldsymbol{\psi}^T}\right\}^{-1} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n) \\
&\qquad - \hat{E}_n\left\{\frac{\partial U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\tau}^T}\right\} \hat{E}_n\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\tau}^T}\right\}^{-1} \\
&\qquad\qquad \times \hat{E}_n\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{\boldsymbol{\tau}}_n) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j, \boldsymbol{O}_i; \hat{\boldsymbol{\tau}}_n) | \boldsymbol{O}_i\}, \\
&\hat{E}_n\{U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) + U(\boldsymbol{O}_j, \boldsymbol{O}_i; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) | \boldsymbol{O}_i\} \\
&\quad = (n-1)^{-1} \sum_{j \neq i} \{U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) + U(\boldsymbol{O}_j, \boldsymbol{O}_i; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)\}, \\
&\hat{E}_n\left\{\frac{\partial U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\psi}^T}\right\} \\
&\quad = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\partial U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\psi}^T}, \\
&\hat{E}_n\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n)}{\partial \boldsymbol{\psi}^T}\right\} = n^{-1} \sum_{i=1}^{n} \frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}_i; \hat{\boldsymbol{\psi}}_n)}{\partial \boldsymbol{\psi}^T}, \\
&\hat{E}_n\left\{\frac{\partial U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\tau}^T}\right\} \\
&\quad = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\partial U(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{v}_{n,\mathrm{DR}}, \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\tau}^T}, \\
&\hat{E}_n\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\tau}^T}\right\} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{\boldsymbol{\tau}}_n)}{\partial \boldsymbol{\tau}^T}, \\
&\hat{E}_n\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{\boldsymbol{\tau}}_n) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j, \boldsymbol{O}_i; \hat{\boldsymbol{\tau}}_n) | \boldsymbol{O}_i\}
\end{aligned}$$

$$= (n-1)^{-1} \sum_{j \neq i} \{ \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_i, \boldsymbol{O}_j; \hat{\boldsymbol{\tau}}_n) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}_j, \boldsymbol{O}_i; \hat{\boldsymbol{\tau}}_n) \}.$$

Inference

Given the estimator $\hat{\sigma}_n^2$ for the asymptotic variance of the doubly robust estimator $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$, an asymptotic $(1-\alpha)100\%$ CI and p-value can be calculated based on the asymptotic normality of the estimator. A $(1-\alpha)100\%$ CI is given by

$$\left[\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) \pm z_{\alpha/2} \hat{\sigma}_n / \sqrt{n} \right]$$

where $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. A p-value for the hypothesis test $H_0 : \nu = \tilde{\nu}$ versus $H_a : \nu \neq \tilde{\nu}$ for some $\tilde{\nu} \in (0,1)$ can be calculated as

$$p = 2 \left\{ 1 - \Phi \left(\left| \frac{\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n) - \tilde{\nu}}{\hat{\sigma}_n / \sqrt{n}} \right| \right) \right\}.$$

7

7.6 Semiparametric Efficiency

In Section 7.4, we demonstrated how we can obtain a doubly robust estimator $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ of the MPI ν_0 using a working model for the propensity score (inducing model $\mathcal{M}(\boldsymbol{\psi})$) and using a working model for the conditional probabilistic index (inducing model $\mathcal{M}(\boldsymbol{\tau})$) and thus $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ is such that it is consistent under the union model $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\tau})$. In Section 7.5, we moreover derived its asymptotic distribution under $\mathcal{M}(\boldsymbol{\psi}) \cup \mathcal{M}(\boldsymbol{\tau})$ and showed how to perform doubly robust inference. In this section, we show that beyond being doubly robust, the estimator $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ is also **locally efficient** within a broad class of estimators: it has smallest asymptotic variance within the class of all estimators that are consistent and asymptotically normal under model $\mathcal{M}(\boldsymbol{\psi})$, provided that also $\mathcal{M}(\boldsymbol{\tau})$ holds. This will follow from a modification of the results presented in Section 6.5.

7.6.1 The space of all influence functions

To obtain the linear variety of all influence functions of CAN estimators of ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$, we will first identify this set assuming the propensity score $\pi_0(\boldsymbol{X})$ is known.

Known propensity score: model $\mathcal{M}_0$

Let $\mathcal{M}_0$ denote the statistical model for the i.i.d. data $\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, defined by the known propensity score $1 > 1-\delta \geq \pi_0(\boldsymbol{X}) \geq \delta > 0$ (with probability one). This is formalized as the set of all density functions

$$\mathcal{M}_0 = \Big\{ f_{Y,A,\boldsymbol{X}}(y,a,\boldsymbol{x};\boldsymbol{\eta}) = f_{Y|A,\boldsymbol{X}}(y|a,\boldsymbol{x};\boldsymbol{\eta}_Y) f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\eta}_{\boldsymbol{X}}) \\ \times \pi_0(\boldsymbol{X})^a \{1-\pi_0(\boldsymbol{X})\}^{(1-a)} : \boldsymbol{\eta} = (\boldsymbol{\eta}_Y, \boldsymbol{\eta}_{\boldsymbol{X}}) \Big\}, \qquad (7.16)$$

with $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_{\boldsymbol{X}}$ infinite dimensional nuisance parameters. It follows from Theorem 2.11 in Chapter 2 that the linear variety of all influence functions is given by $\mathcal{V}_0 = \phi_0 + \mathcal{T}_0^{\perp}$, with $\mathcal{T}_0^{\perp}$ the orthogonal complement of the tangent space $\mathcal{T}_0$ of the statistical model $\mathcal{M}_0$ and ϕ_0 and arbitrary influence function. From the same reasoning as in Section 6.5, it follows that $\mathcal{T}_0 = \mathcal{T}_Y \oplus \mathcal{T}_{\boldsymbol{X}}$, with $\mathcal{T}_Y = \big\{ \alpha_Y(Y,A,\boldsymbol{X}) : E\{\alpha_Y(Y,A,\boldsymbol{X})|A,\boldsymbol{X}\} = 0 \big\}$ (the tangent space corresponding to the conditional density function $f_{Y|A,\boldsymbol{X}}(y|a,\boldsymbol{x};\boldsymbol{\eta}_Y)$) and $\mathcal{T}_{\boldsymbol{X}} = \big\{ \alpha_{\boldsymbol{X}}(\boldsymbol{X}) : E\{\alpha_{\boldsymbol{X}}(\boldsymbol{X})\} = 0 \big\}$ (the tangent space corresponding to the marginal density function $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\eta}_{\boldsymbol{X}})$) with both $\alpha_Y(Y,A,\boldsymbol{X})$ and $\alpha_{\boldsymbol{X}}(\boldsymbol{X})$ square-integrable. Note that $\mathcal{T}_Y \perp \mathcal{T}_{\boldsymbol{X}}$. It then follows from Theorem 2.12 that $\mathcal{T}_0^{\perp} = \big\{ \alpha(A,\boldsymbol{X}) : E\{\alpha(A,\boldsymbol{X})|\boldsymbol{X}\} = 0 \big\}$ with $\alpha(A,\boldsymbol{X})$ square-integrable, which can be equivalently written as $\mathcal{T}_0^{\perp} = \big\{ \{A-\pi_0(\boldsymbol{X})\}\tilde{\alpha}(\boldsymbol{X}) : \tilde{\alpha}(\boldsymbol{X})$ arbitrary square-integrable function of $\boldsymbol{X} \big\}$, which equals the tangent space $\mathcal{T}_A$, corresponding to an unspecified propensity score. It follows that $\mathcal{V}_0 = \big\{ \phi_0 + \{A-\pi_0(\boldsymbol{X})\}\tilde{\alpha}(\boldsymbol{X}) \big\}$ with ϕ_0 an arbitrary influence function of ν_0 under the statistical model $\mathcal{M}_0$. E.g., we may choose the influence function of the IPTW estimator with known propensity score:

$$\hat{\nu}^{(0)}_{n,\text{IPTW}} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{(1-A_i)A_j I_{ij}}{\{1-\pi_0(\boldsymbol{X}_i)\}\pi_0(\boldsymbol{X}_j)}$$

with $I_{ij} = I(Y_i \preceq Y_j)$. This can be obtained via the Hájek projection (Hájek 1970; van der Vaart 1998) of $\hat{\nu}^{(0)}_{n,\text{IPTW}}$, which is given by

$$\begin{aligned}\tilde{\nu}^{(0)}_{n,\text{IPTW}} &= \sum_{i=1}^{n} E(\hat{\nu}^{(0)}_{n,\text{IPTW}} | Y_i, A_i, \boldsymbol{X}_i) - \nu_0 \\ &= n^{-1} \sum_{i=1}^{n} \frac{1-A_i}{1-\pi_0(\boldsymbol{X}_i)} a_1^{(0)}(Y_i) - \nu_0 + \frac{A_i}{\pi_0(\boldsymbol{X}_i)} a_2^{(0)}(Y_i) - \nu_0,\end{aligned}$$

with $a_1^{(0)}(Y_i) = E[\{A_j/\pi_0(\boldsymbol{X}_j)\}I_{ij}|Y_i]$ and $a_2^{(0)}(Y_i) = E[(1-A_j)/\{1-\pi(\boldsymbol{X}_j)\}I_{ji}|Y_i]$. Because $n^{1/2}(\hat{\nu}^{(0)}_{n,\text{IPTW}} - \tilde{\nu}^{(0)}_{n,\text{IPTW}}) = o_p(1)$ (see Theorem 12.3, van der Vaart (1998)), it follows that $n^{1/2}(\hat{\nu}^{(0)}_{n,\text{IPTW}} - \nu_0) = n^{-1/2}\sum_{i=1}^{n} \phi_0(Y_i, A_i, \boldsymbol{X}_i; \nu_0) + o_p(1)$ and thus is asymptotically linear with influence function $\phi_0(Y_i, A_i, \boldsymbol{X}_i; \nu_0) = [(1-A_i)/\{1-\pi_0(\boldsymbol{X}_i)\}]a_1^{(0)}(Y_i) - \nu_0 + \{A_i/\pi_0(\boldsymbol{X}_i)\}a_2^{(0)}(Y_i) - \nu_0$. We thus have proven the following result:

Theorem 7.3 (Space of Influence Functions under $\mathcal{M}_0$)**.** *The space of all influence functions of ν_0 under model $\mathcal{M}_0$ is given by*

$$\mathcal{V}_0 = \left\{ \phi^{(0)}_{\tilde{\alpha}}(Y, A, \boldsymbol{X}; \nu_0) = \frac{1-A}{1-\pi_0(\boldsymbol{X})} a_1^{(0)}(Y) - \nu_0 + \frac{A}{\pi_0(\boldsymbol{X})} a_2^{(0)}(Y) - \nu_0 + \{A - \pi_0(\boldsymbol{X})\}\tilde{\alpha}(\boldsymbol{X}) : \right.$$
$$\left. \tilde{\alpha}(\boldsymbol{X}) \text{ an arbitrary square-integrable function of } \boldsymbol{X} \right\}.$$

7

Unknown but correctly specified propensity score: model $\mathcal{M}(\boldsymbol{\psi})$

Building on the results of Theorem 7.3, we will now derive the linear variety $\mathcal{V}$ of influence functions for ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$, the statistical model induced by a parametric working model $\pi(\boldsymbol{X}; \boldsymbol{\psi})$ for the propensity score and $\boldsymbol{\psi}_0$ well-defined as $\pi(\boldsymbol{X}; \boldsymbol{\psi}_0) = \pi_0(\boldsymbol{X})$ (where we again assume positivity). This can be formalized as the set of all joint density functions

$$\mathcal{M}(\boldsymbol{\psi}) = \left\{ f_{Y,A,\boldsymbol{X}}(y, a, \boldsymbol{x}; \boldsymbol{\psi}, \boldsymbol{\eta}) = f_{Y|A,\boldsymbol{X}}(y|a, \boldsymbol{x}; \boldsymbol{\eta}_Y) f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\eta}_{\boldsymbol{X}}) \right.$$

$$\times \pi(\boldsymbol{x};\boldsymbol{\psi})^a\{1-\pi(\boldsymbol{x};\boldsymbol{\psi})\}^{(1-a)} : \boldsymbol{\psi}, \boldsymbol{\eta} = (\boldsymbol{\eta}_Y, \boldsymbol{\eta}_{\boldsymbol{X}})\Big\}, \qquad (7.17)$$

with $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_{\boldsymbol{X}}$ infinite dimensional nuisance parameters and $\boldsymbol{\psi}$ an s-dimensional nuisance parameter. The tangent space of $\mathcal{M}(\boldsymbol{\psi})$ is then given by the direct sum $\mathcal{T} = \mathcal{T}_0 \oplus \Lambda_{\boldsymbol{\psi}}$, where $\Lambda_{\boldsymbol{\psi}} = \{\boldsymbol{b}^T \boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0) | \boldsymbol{b} \in \mathbb{R}^s\}$ with $\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0) = \{A-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}/[\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}]\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)$ the score for $\boldsymbol{\psi}$, and with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0) = \partial\pi(\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$. Because $\Lambda_{\boldsymbol{\psi}} \subset \mathcal{T}_0^{\perp}$, $\Lambda_{\boldsymbol{\psi}} \perp \mathcal{T}_0$. An arbitrary influence function $\phi_{\tilde{\alpha}}$ for ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$ can hence be obtained as the residual after projecting any $\phi_{\tilde{\alpha}}^{(0)} \in \mathcal{V}_0$ onto $\Lambda_{\boldsymbol{\psi}}$; that is, $\phi_{\tilde{\alpha}} = \phi_{\tilde{\alpha}}^{(0)} - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\Lambda_{\boldsymbol{\psi}}) = \phi_{\tilde{\alpha}}^{(0)} - E(\phi_{\tilde{\alpha}}^{(0)}\boldsymbol{S}_{\boldsymbol{\psi}}^T)E^{-1}(\boldsymbol{S}_{\boldsymbol{\psi}}\boldsymbol{S}_{\boldsymbol{\psi}}^T)\boldsymbol{S}_{\boldsymbol{\psi}}$. We find that

$$\begin{aligned} E(\phi_{\tilde{\alpha}}^{(0)}\boldsymbol{S}_{\boldsymbol{\psi}}^T) &= E\Bigg(\left[\frac{E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \frac{E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} + \tilde{\alpha}(\boldsymbol{X})\right] \\ &\qquad \times \pi_{\boldsymbol{\psi}}^T(\boldsymbol{X};\boldsymbol{\psi}_0)\Bigg), \\ E(\boldsymbol{S}_{\boldsymbol{\psi}}\boldsymbol{S}_{\boldsymbol{\psi}}^T) &= E\left[\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\pi_{\boldsymbol{\psi}}^T(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}}\right]. \end{aligned}$$

We may conclude with the following result:

Theorem 7.4 (Space of Influence Function under $\mathcal{M}(\boldsymbol{\psi})$)**.** *The space of all influence functions of ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$ is given by*

$$\begin{aligned} \mathcal{V} = \Bigg\{ & \phi_{\tilde{\alpha}}(Y,A,\boldsymbol{X};\nu_0) = \frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}a_1^{(0)}(Y) - \nu_0 + \frac{A}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}a_2^{(0)}(Y) - \nu_0 \\ & \qquad + \{A-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}\tilde{\alpha}_{\boldsymbol{\psi}}(\boldsymbol{X}) : \\ & \tilde{\alpha}_{\boldsymbol{\psi}}(\boldsymbol{X}) = \tilde{\alpha}(\boldsymbol{X}) - E(\phi_{\tilde{\alpha}}^{(0)}\boldsymbol{S}_{\boldsymbol{\psi}}^T)E(\boldsymbol{S}_{\boldsymbol{\psi}}\boldsymbol{S}_{\boldsymbol{\psi}}^T)\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}} \\ & \textit{and } \tilde{\alpha}(\boldsymbol{X}) \textit{ an arbitrary square-integrable function of } \boldsymbol{X} \Bigg\}. \end{aligned}$$

7.6.2 The efficient influence function

Theorem 7.4 provides the class $\mathcal{V}$ of all influence functions $\phi_{\tilde{\alpha}}$ of ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$. It now remains to identify the function $\tilde{\alpha}_{\text{eff}}(\boldsymbol{X})$ leading to the influence function $\phi_{\text{eff}}(Y,A,\boldsymbol{X};\nu_0) \equiv \phi_{\tilde{\alpha}_{\text{eff}}}(Y,A,\boldsymbol{X};\nu_0)$ that has smallest variance among all influence functions $\phi_{\tilde{\alpha}}(Y,A,\boldsymbol{X};\nu_0)$, which is the content of the following theorem.

Theorem 7.5 (Efficient Influence Function)**.** *The choice*

$$\tilde{\alpha}_{\mathit{eff}}(\boldsymbol{X}) = \left[\frac{E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \frac{E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right] \tag{7.18}$$

delivers the efficient influence function of ν_0 *under model* $\mathcal{M}(\boldsymbol{\psi})$*:*

$$\begin{aligned}
&\phi_{\mathit{eff}}(Y,A,\boldsymbol{X};\nu_0) \qquad (7.19)\\
&= \frac{1-A}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}a_1^{(0)}(Y) - \nu_0 + \frac{A}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}a_2^{(0)}(Y) - \nu_0\\
&\quad + \{A-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)\}\left[\frac{E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}}{1-\pi(\boldsymbol{X};\boldsymbol{\psi}_0)} - \frac{E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}}{\pi(\boldsymbol{X};\boldsymbol{\psi}_0)}\right].
\end{aligned}$$

Proof. An arbitrary element of $\mathcal{V}$, the linear variety of influence functions under model $\mathcal{M}(\boldsymbol{\psi})$, can be written as $\phi_{\tilde{\alpha}} = \phi_{\tilde{\alpha}}^{(0)} - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\Lambda_{\boldsymbol{\psi}})$, with $\phi_{\tilde{\alpha}}^{(0)}$ an arbitrary element of $\mathcal{V}_0$, the linear variety of influence functions under model $\mathcal{M}_0$. If follows from Theorem 2.11 that the efficient influence function for ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$ is given by

$$\phi_{\text{eff}} = \phi_{\tilde{\alpha}} - \Pi(\phi_{\tilde{\alpha}}|\mathcal{T}^{\perp}).$$

Because $\mathcal{T} = \mathcal{T}_0 \oplus \Lambda_{\boldsymbol{\psi}}$, $\Lambda_{\boldsymbol{\psi}} \subset \mathcal{T}_A$, it follows that $\mathcal{T}^{\perp} = \mathcal{T}_A \cap \Lambda_{\boldsymbol{\psi}}^{\perp}$ and thus that $\mathcal{T}^{\perp} = \Pi(\mathcal{T}_A|\Lambda_{\boldsymbol{\psi}}^{\perp}) = \mathcal{T}_A - \Pi(\mathcal{T}_A|\Lambda_{\boldsymbol{\psi}})$. From this, we find that

$$\begin{aligned}
\phi_{\text{eff}} &= \phi_{\tilde{\alpha}} - \Pi(\phi_{\tilde{\alpha}}|\mathcal{T}^{\perp})\\
&= \phi_{\tilde{\alpha}} - \Pi(\phi_{\tilde{\alpha}}|\mathcal{T}_A) + \Pi\{\Pi(\phi_{\tilde{\alpha}}|\mathcal{T}_A)|\Lambda_{\boldsymbol{\psi}}\}\\
&= \phi_{\tilde{\alpha}}^{(0)} - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\Lambda_{\boldsymbol{\psi}}) - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\mathcal{T}_A) + \Pi\{\Pi(\phi_{\tilde{\alpha}}^{(0)}|\Lambda_{\boldsymbol{\psi}})|\mathcal{T}_A\} + \Pi(\phi_{\tilde{\alpha}}|\Lambda_{\boldsymbol{\psi}})\\
&= \phi_{\tilde{\alpha}}^{(0)} - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\Lambda_{\boldsymbol{\psi}}) - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\mathcal{T}_A) + \Pi(\phi_{\tilde{\alpha}}^{(0)}|\Lambda_{\boldsymbol{\psi}})
\end{aligned}$$

$$= \phi_{\tilde{\alpha}}^{(0)} - \Pi(\phi_{\tilde{\alpha}}^{(0)}|\mathcal{T}_A),$$

which, by Theorem 2.11, equals the efficient influence function for ν_0 under model $\mathcal{M}_0$, and thus $\phi_{\text{eff}} = \Pi(\phi_0|\mathcal{T}_0) = \phi_0 - E(\phi_0|A,\boldsymbol{X}) + E(\phi_0|X) = \phi_0 + \{A - \pi_0(\boldsymbol{X})\}\tilde{\alpha}_{\text{eff}}(\boldsymbol{X})$, with $\tilde{\alpha}_{\text{eff}}(\boldsymbol{X})$ given in (7.18). □

7.6.3 Local efficiency of $\hat{\nu}_{n,\mathbf{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$

Theorem 7.5 gives an expression of the efficient influence function of ν_0 under the semiparametric model that assumes a correctly specified propensity score model, model $\mathcal{M}(\boldsymbol{\psi})$. We next argue that under correct specification of $\mathcal{M}(\boldsymbol{\psi})$, the doubly robust estimator $\hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ has influence function in the class $\mathcal{V}$ (provided $\hat{\boldsymbol{\psi}}_n$ is estimated via MLE) and show that under the intersection model $\mathcal{M}(\boldsymbol{\psi}) \cap \mathcal{M}(\boldsymbol{\tau})$ (for arbitrary root-n consistent nuisance parameter estimators), $\hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ has influence function equal to the efficient influence function; that is, $\phi_\nu(\boldsymbol{O}; \nu_0, \boldsymbol{\psi}_0, \boldsymbol{\tau}_0) = \phi_{\text{eff}}(Y, A, \boldsymbol{X}; \nu_0)$, making it locally efficient.

Recall that for working models $\pi(\boldsymbol{X}; \boldsymbol{\psi})$ and $m(A, A^*, \boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\tau})$ and for root-n consistent estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\tau}}_n$ with probability limits $\boldsymbol{\psi}^*$ and $\boldsymbol{\tau}^*$ (assuming that at least on of both working models is correctly specified), the influence function $\phi_\nu(\boldsymbol{O}; \nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)$ of the doubly robust estimator $\hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ is given by

$$\begin{aligned}
&E\{U(\boldsymbol{O}, \boldsymbol{O}^*; \nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*) + U(\boldsymbol{O}^*, \boldsymbol{O}; \nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)|\boldsymbol{O}\} \\
&- E\left\{\frac{\partial U(\boldsymbol{O}, \boldsymbol{O}^*; \nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)}{\partial \boldsymbol{\psi}^T}\right\} E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}; \boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^T}\right\}^{-1} \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O}; \boldsymbol{\psi}^*) \\
&- E\left\{\frac{\partial U(\boldsymbol{O}, \boldsymbol{O}^*; \nu_0, \boldsymbol{\psi}^*, \boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\} E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}, \boldsymbol{O}^*; \boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}^T}\right\}^{-1} \\
&\qquad \times E\{\boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}, \boldsymbol{O}^*; \boldsymbol{\tau}^*) + \boldsymbol{U}_{\boldsymbol{\tau}}(\boldsymbol{O}^*, \boldsymbol{O}; \boldsymbol{\tau}^*)|\boldsymbol{O}\}
\end{aligned}$$

with the estimating function

$$\begin{aligned}
U(\boldsymbol{O}, \boldsymbol{O}^*; \nu, \boldsymbol{\psi}, \boldsymbol{\tau}) &= m(0, 1, \boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\tau}) - \nu \\
&\quad + \frac{1 - A}{1 - \pi(\boldsymbol{X}; \boldsymbol{\psi})} \frac{A^*}{\pi(\boldsymbol{X}^*; \boldsymbol{\psi})} \{I(Y \preceq Y^*) - m(0, 1, \boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\tau})\}.
\end{aligned}$$

Influence function of $\hat{v}_{n,\mathbf{DR}}(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\tau}}_n)$ under model $\mathcal{M}(\boldsymbol{\psi})$

Assuming that the propensity score working model is correctly specified, in which case $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, it follows from an easy calculation that

$$\begin{aligned}
&E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)+U(\boldsymbol{O}^*,\boldsymbol{O};\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)|\boldsymbol{O}\} \\
&\quad = \frac{1-A}{1-\pi_0(\boldsymbol{X})}a_1^{(0)}(Y)-\nu_0+\frac{A}{\pi_0(\boldsymbol{X})}a_2^{(0)}(Y)-\nu_0 \\
&\qquad +\{A-\pi_0(\boldsymbol{X})\}\left[\frac{E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}-\frac{E\{m(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\right] \\
&\quad = \phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\in\mathcal{V}_0
\end{aligned}$$

with

$$\tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X})=\left[\frac{E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}-\frac{E\{m(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\right]. \tag{7.20}$$

When $\hat{\boldsymbol{\psi}}_n$ is obtained via MLE, so that $\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}_0)=\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)$, if follows that

$$\begin{aligned}
&E^{-1}\{\partial\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}_0)/\partial\boldsymbol{\psi}^T\}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}_0) \\
&\qquad = -E^{-1}\{\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)\boldsymbol{S}^T_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)\}\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0).
\end{aligned}$$

Next observe that

$$\begin{aligned}
&E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)}{\partial\boldsymbol{\psi}}\right\} \\
&\quad = E\Bigg[(1-A)A^*\{I(Y\preceq Y^*)-m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\} \\
&\qquad\qquad \times\frac{\partial}{\partial\boldsymbol{\psi}}\left\{\frac{1}{\{1-\pi(\boldsymbol{X};\boldsymbol{\psi})\}\pi(\boldsymbol{X}^*;\boldsymbol{\psi})}\right\}\Bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\Bigg] \\
&\quad = E\Bigg[\{P(Y\preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*)-m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\} \\
&\qquad\qquad \times\left\{\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{1-\pi_0(\boldsymbol{X})}-\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X}^*;\boldsymbol{\psi}_0)}{\pi_0(\boldsymbol{X}^*)}\right\}\Bigg]
\end{aligned}$$

$$= E\left[\{P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*) - m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\}\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{1-\pi_0(\boldsymbol{X})}\right]$$
$$- E\left[\{P(Y^* \preceq Y|A^*=0,A=1,\boldsymbol{X}^*,\boldsymbol{X}) - m(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*)\}\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{\pi_0(\boldsymbol{X})}\right].$$

From the law of iterated expectation, it then follows that

$$E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)}{\partial\boldsymbol{\psi}}\right\}$$
$$= E\left[\frac{E\{P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*) - m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\right]$$
$$- E\left[\frac{E\{P(Y^* \preceq Y|A^*=0,A=1,\boldsymbol{X}^*,\boldsymbol{X}) - m(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\right].$$

Next, because

$$\begin{aligned} E\{P(Y \preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*)|\boldsymbol{X}\} &= E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}, \\ E\{P(Y^* \preceq Y|A^*=0,A=1,\boldsymbol{X}^*,\boldsymbol{X})|\boldsymbol{X}\} &= E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}, \end{aligned}$$

it follows that

$$E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)}{\partial\boldsymbol{\psi}}\right\}$$
$$= -E\left\{\left(\left[\frac{E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})} - \frac{E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}\right]\right.\right.$$
$$\left.\left. + \left[\frac{E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})} - \frac{E\{m(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\right]\right)\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\right\}$$
$$= -E\left(\left[\frac{E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})} - \frac{E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})} + \tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X})\right]\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)\right),$$

showing that

$$E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)}{\partial\boldsymbol{\psi}}\right\} = -E\{\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)\}.$$

From these calculations, it follows that

$$E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\psi}^T}\right\}E\left\{\frac{\partial \boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^T}\right\}^{-1}\boldsymbol{U}_{\boldsymbol{\psi}}(\boldsymbol{O};\boldsymbol{\psi}_0)$$
$$= E\{\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\boldsymbol{S}^T_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)\}$$
$$\times E^{-1}\{\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)\boldsymbol{S}^T_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0)\}\boldsymbol{S}_{\boldsymbol{\psi}}(A,\boldsymbol{X};\boldsymbol{\psi}_0).$$

Finally, note that under model $\mathcal{M}(\boldsymbol{\psi})$,

$$E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)}{\partial \boldsymbol{\tau}}\right\}$$
$$= E\left[\left\{1-\frac{1-A}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}\right\}m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\right]$$
$$= E\left[\left\{1-\frac{E(1-A|\boldsymbol{X})}{1-\pi_0(\boldsymbol{X})}\frac{E(A^*|\boldsymbol{X}^*)}{\pi_0(\boldsymbol{X}^*)}\right\}m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)\right]=\boldsymbol{0},$$

with $m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)=\partial m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\boldsymbol{\tau}^*}$. We can conclude with the following results:

7

Proposition 7.3 (Influence function of $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ under $\mathcal{M}(\boldsymbol{\psi})$)**.** *Under model $\mathcal{M}(\boldsymbol{\psi})$, the influence function of the doubly robust estimator $\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$, is given by*

$$\phi_{\nu}(\boldsymbol{O};\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}^*)=\phi_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\in\mathcal{V},$$

provided that $\hat{\boldsymbol{\psi}}_n$ is obtained via MLE and with $\tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X})$ given by equation (7.20).

Influence function of $\hat{\nu}_{n,\mathbf{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ under model $\mathcal{M}(\boldsymbol{\psi})\cap\mathcal{M}(\boldsymbol{\tau})$

Under the intersection model $\mathcal{M}(\boldsymbol{\psi})\cap\mathcal{M}(\boldsymbol{\tau})$, not only $\boldsymbol{\psi}^*=\boldsymbol{\psi}_0$, but also $\boldsymbol{\tau}^*=\boldsymbol{\tau}_0$, so that $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}_0)=m_0(A,A^*,\boldsymbol{X},\boldsymbol{X}^*)=P(Y\preceq Y^*|A,A^*,\boldsymbol{X},\boldsymbol{X}^*)$. We already have seen that $E\{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}_0)/\partial\boldsymbol{\tau}\}=\boldsymbol{0}$. We now additionally

have that

$$
\begin{aligned}
&E\left\{\frac{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}_0)}{\partial\boldsymbol{\psi}}\right\}\\
&=E\Bigg[\{P(Y\preceq Y^*|A=0,A^*=1,\boldsymbol{X},\boldsymbol{X}^*)-m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}_0)\}\\
&\qquad\qquad\times\left\{\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\boldsymbol{\psi}_0)}{1-\pi_0(\boldsymbol{X})}-\frac{\pi_{\boldsymbol{\psi}}(\boldsymbol{X}^*;\boldsymbol{\psi}_0)}{\pi_0(\boldsymbol{X}^*)}\right\}\Bigg]=\mathbf{0}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\phi_\nu(\boldsymbol{O};\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}_0)\\
&\quad=E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}_0)+U(\boldsymbol{O}^*,\boldsymbol{O};\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}_0)|\boldsymbol{O}\}\\
&\quad=\frac{1-A}{\pi_0(\boldsymbol{X})}a_1^{(0)}(Y)-\nu_0+\frac{A}{\pi_0(\boldsymbol{X})}a_2^{(0)}(Y)-\nu_0\\
&\qquad+\{A-\pi_0(\boldsymbol{X})\}\left[\frac{E\{m_0(0,1,\boldsymbol{X},\boldsymbol{X}^*)|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}-\frac{E\{m_0(0,1,\boldsymbol{X}^*,\boldsymbol{X})|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\right]\\
&\quad=\frac{1-A}{\pi_0(\boldsymbol{X})}a_1^{(0)}(Y)-\nu_0+\frac{A}{\pi_0(\boldsymbol{X})}a_2^{(0)}(Y)-\nu_0\\
&\qquad+\{A-\pi_0(\boldsymbol{X})\}\left[\frac{E\{a_1^{(0)}(Y)|A=0,\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}-\frac{E\{a_2^{(0)}(Y)|A=1,\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\right].
\end{aligned}
$$

We thus have the following:

Proposition 7.4 (Influence function of $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ under $\mathcal{M}(\boldsymbol{\psi})\cap\mathcal{M}(\boldsymbol{\tau})$)**.** *Under the intersection model $\mathcal{M}(\boldsymbol{\psi})\cap\mathcal{M}(\boldsymbol{\tau})$, the influence function of the doubly robust estimator $\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$, is given by*

$$\phi_\nu(\boldsymbol{O};\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau}_0)=\phi_{eff}(Y,A,\boldsymbol{X};\nu_0),$$

for arbitrary root-n consistent nuisance parameter estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\tau}}_n$.

We may hence conclude that $\hat{\nu}_{n,\mathrm{DR}}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$ is a locally efficient estimator of ν_0 under model $\mathcal{M}(\boldsymbol{\psi})$ at $\mathcal{M}(\boldsymbol{\tau})$.

Remark 7.2. *The property that under model $\mathcal{M}(\boldsymbol{\psi})$, the gradient $E\{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_0,\boldsymbol{\tau})/\partial\boldsymbol{\tau}\}=\mathbf{0}$ for all $\boldsymbol{\tau}$ and that under model $\mathcal{M}(\boldsymbol{\tau})$, the gradient $E\{\partial U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi},\boldsymbol{\tau}_0)/\partial\boldsymbol{\psi}\}=\mathbf{0}$ for all $\boldsymbol{\psi}$ is of no coincidence and follows from the double robustness of the estimator $\hat{\nu}_{n,DR}(\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\tau}}_n)$, see also Section 3.4.*

7.7 Discussion

In this chapter, we have presented a doubly robust adaptation of the Mann-Whitney U test statistic as an estimator for the MPI $\nu_0 = P\{Y(0) \preceq Y^*(1)\}$ in the presence of confounding for the treatment-outcome associations. The resulting estimator consistently estimates the MPI if we either correctly specify a working model for the propensity score (the conditional distribution of treatment given confounders) or a working model for the conditional probabilistic index of the outcome given the treatment and confounders. We additionally demonstrated how to obtain doubly robust standard errors of the doubly robust estimator and how to construct a Wald test for the hypothesis test $H_0 : \nu = \tilde{\nu}$ versus $H_a : \nu \neq \tilde{\nu}$ for some $\tilde{\nu} \in (0,1)$. This generalizes the use of the Mann-Whitney U test, which is confined to the analysis of randomized experiments if interest lies in the assessment of a causal effect, to the analysis of observational studies. We moreover showed that the doubly robust estimator is locally efficient in the sense that it has smallest asymptotic variance within the class of all estimators that are consistent and asymptotically normal under a correctly specified propensity score model, provided that the CPI working model is also correctly specified.

As seen in earlier chapters, the actual benefit of using a doubly robust estimator of ν_0 could be questioned in practice because then, it is likely that misspecification affects both working models. It thus remains to empirically evaluate the advantages (both in terms of bias and efficiency) of the doubly robust estimator as compared to simpler estimators, such as the regression imputation estimator of Section 7.3.1 and the IPTW estimator of Section 7.3.2, as well as compared to the alternatives of Wu et al. (2013) and Chen et al. (2013). We plan to evaluate this in future research.

A further question is how to optimally choose estimators $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\tau}}_n$ of the

nuisance parameters $\boldsymbol{\psi}$ and $\boldsymbol{\tau}$ indexing the working models $\mathcal{M}(\boldsymbol{\psi})$ and $\mathcal{M}(\boldsymbol{\tau})$. This question arises because under misspecification of at least one working model, the asymptotic distribution of the doubly robust estimator depends on the choice of nuisance parameter estimators, as seen in Section 7.5 (see also Section 3.4). A convenient choice would be to choose the MLE $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ of $\boldsymbol{\psi}$ and to use the strategy outlined in Appendix 6.A, proposed in Thas et al. (2012), to obtain an estimator $\hat{\boldsymbol{\tau}}_n^{\text{T}}$ of $\boldsymbol{\tau}$, where T indicates Thas. This would result in the doubly robust estimator $\hat{\nu}_{n,\text{DR}}^{\text{T}} \equiv \hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}, \hat{\boldsymbol{\tau}}_n^{\text{T}})$. However, potentially more clever choices could be made. Below, we list several possibilities.

Doubly robust regression imputation estimator

A first option would be to construct a doubly robust regression imputation estimator. In the TMLE-literature, this is also referred to as a doubly robust substitution estimator. For this purpose, we use the MLE $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$ as an estimator of $\boldsymbol{\psi}$. The parameter $\boldsymbol{\tau}$ indexing the PIM $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}) = \text{expit}\{\tau_A(A^*-A) + \boldsymbol{\tau}_{\boldsymbol{X}}^T(\boldsymbol{X}^* - \boldsymbol{X})\}$, with $\text{expit}(x) = 1/(1+e^{-x})$, is then estimated via a weighted regression; that is, we let $\hat{\boldsymbol{\tau}}_n^{\text{WR}}$ denote the solution to the estimating equation

$$\mathbf{0} = \sum_{i=1}^{n}\sum_{j\neq i} W^{(1)}(A_i,\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})W^{(1)}(A_j,\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{MLE}}) \times \{I_{ij} - m(A_i,A_j,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n^{\text{WR}})\}\begin{bmatrix} A_j - A_i \\ \boldsymbol{X}_j - \boldsymbol{X}_i \end{bmatrix},$$

where $I_{ij} = I(Y_i \preceq Y_j)$ and $W^{(1)}(A,\boldsymbol{X},\boldsymbol{\psi}) = [\pi(\boldsymbol{X};\boldsymbol{\psi})^A\{1-\pi(\boldsymbol{X};\boldsymbol{\psi})\}^{1-A}]^{-1}$. From a similar reasoning as on page 195, this estimating equation implies that

$$0 = \sum_{i=1}^{n}\sum_{j\neq i}\frac{1-A_i}{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}\frac{A_j}{\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{MLE}})}\{I_{ij} - m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n^{\text{WR}})\},$$

showing that

$$\hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}},\hat{\boldsymbol{\tau}}_n^{\text{WR}}) = \{n(n-1)\}^{-1}\sum_{i=1}^{n}\sum_{j\neq i} m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n^{\text{WR}}),$$

resulting in a doubly robust regression imputation estimator obtained by standardizing PIM predictions based on the estimator $\hat{\boldsymbol{\tau}}_n^{\text{WR}}$, see also Section 6.4. The advantage of this doubly robust estimator $\hat{\nu}_{n,\text{DR}}^{\text{WR}} \equiv \hat{\nu}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{MLE}}, \hat{\boldsymbol{\tau}}_n^{\text{WR}})$ is that it guarantees to give estimates that fall within the allowed parameter range, that is, between zero and one.

Bias-reduced doubly robust estimator

A second alternative would be to exploit the bias-reduced doubly robust estimation principle, developed in Chapter 4. For working models $\pi(\boldsymbol{X};\boldsymbol{\psi})$ and $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$, the first-order asymptotic bias of the doubly robust estimator is given by the mean of the influence function as if the nuisance parameters would be known: $\text{bias}(\boldsymbol{\psi},\boldsymbol{\tau};\nu_0) = E[E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi},\boldsymbol{\tau}) + U(\boldsymbol{O}^*,\boldsymbol{O};\nu_0,\boldsymbol{\psi},\boldsymbol{\tau})|\boldsymbol{O}\}]$ and by Theorem 4.1, it follows that the squared first-order asymptotic bias of the doubly robust estimator is locally minimized in the values $(\boldsymbol{\psi}_{\text{BR}}^{*,T}, \boldsymbol{\tau}_{\text{BR}}^{*,T})^T$, defined as the solutions to the equations $E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_{\text{BR}}^*,\boldsymbol{\tau}_{\text{BR}}^*)/\partial\boldsymbol{\tau}\} = \boldsymbol{0}$ and $E\{U(\boldsymbol{O},\boldsymbol{O}^*;\nu_0,\boldsymbol{\psi}_{\text{BR}}^*,\boldsymbol{\tau}_{\text{BR}}^*)/\partial\boldsymbol{\psi}\} = \boldsymbol{0}$. The bias-reduced estimation principle would thus lead to nuisance parameter estimators $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\tau}}_n^{\text{BR}}$ that solve the estimating equations

$$\boldsymbol{0} = \sum_{i=1}^{n}\sum_{j\neq i} \partial U(\boldsymbol{O}_i,\boldsymbol{O}_j;\nu_0,\hat{\boldsymbol{\psi}}_n^{\text{BR}},\hat{\boldsymbol{\tau}}_n^{\text{BR}})/\partial\boldsymbol{\tau},$$
$$\boldsymbol{0} = \sum_{i=1}^{n}\sum_{j\neq i} \partial U(\boldsymbol{O}_i,\boldsymbol{O}_j;\nu_0,\hat{\boldsymbol{\psi}}_n^{\text{BR}},\hat{\boldsymbol{\tau}}_n^{\text{BR}})/\partial\boldsymbol{\psi},$$

which deliver consistent nuisance parameter estimators under correct working model specification (see Theorem 4.2). Specifically, $\hat{\boldsymbol{\psi}}_n^{\text{BR}}$ and $\hat{\boldsymbol{\tau}}_n^{\text{BR}}$ are solutions to the estimating equations

$$\boldsymbol{0} = \sum_{i=1}^{n}\sum_{j\neq i} \left\{1 - \frac{1-A_i}{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{BR}})}\frac{A_j}{\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{BR}})}\right\} m_{\boldsymbol{\tau}}(0,1;\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n^{\text{BR}}),$$
$$\boldsymbol{0} = \sum_{i=1}^{n}\sum_{j\neq i} (1-A_i)A_j\{I_{ij} - m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n^{\text{BR}})\}W^{(2)}(\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{BR}}),$$

where $I_{ij} = I(Y_i \preceq Y_j)$, $m_{\boldsymbol{\tau}}(0,1;\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}_n^{\text{BR}}) = \partial m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\tau})/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\hat{\boldsymbol{\tau}}_n^{\text{BR}}}$, and with weights

$$W^{(2)}(\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{BR}}) = \frac{\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{BR}})\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{BR}}) - \{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{BR}})\}\pi_{\boldsymbol{\psi}}(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{BR}})}{[\{1-\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\psi}}_n^{\text{BR}})\}\pi(\boldsymbol{X}_j;\hat{\boldsymbol{\psi}}_n^{\text{BR}})]^2}$$

with $\pi_{\boldsymbol{\psi}}(\boldsymbol{X};\hat{\boldsymbol{\psi}}_n^{\text{BR}}) = \partial\pi(\boldsymbol{X};\boldsymbol{\psi})/\partial\boldsymbol{\psi}|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_n^{\text{BR}}}$. The bias-reduced doubly robust estimator is then given by $\hat{v}_{n,\text{DR}}^{\text{BR}} \equiv \hat{v}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}_n^{\text{BR}},\hat{\boldsymbol{\tau}}_n^{\text{BR}})$. Aside from the bias-reduction property, this estimation principle also eliminates the estimator's first-order dependence on the nuisance parameter estimators in the sense that the doubly robust estimator is first-order ancillary with respect to the nuisance parameters (see Corollary 4.1). This simplifies standard error calculation, which can now be easily obtained as

$$\frac{\left[\sum_{i=1}^n\sum_{j\neq i}\left\{U_{ij}(\hat{v}_{n,\text{DR}}^{\text{BR}},\hat{\boldsymbol{\psi}}_n^{\text{BR}},\hat{\boldsymbol{\tau}}_n^{\text{BR}}) + U_{ji}(\hat{v}_{n,\text{DR}}^{\text{BR}},\hat{\boldsymbol{\psi}}_n^{\text{BR}},\hat{\boldsymbol{\tau}}_n^{\text{BR}})\right\}^2\right]^{1/2}}{\{n^2(n-1)\}^{1/2}},$$

with $U_{ij}(v,\boldsymbol{\psi},\boldsymbol{\tau}) = U(\boldsymbol{O}_i,\boldsymbol{O}_j;v,\boldsymbol{\psi},\boldsymbol{\tau})$. However, it remains to be evaluated how to best solve these estimating equations for the nuisance parameters and moreover, if it could be worthwhile to exploit bias-reduction in a single direction only (see Chapter 5).

Empirical efficiency maximization

A third possibility would be to exploit the idea behind empirical efficiency maximization (EEM), originally proposed in Rubin and van der Laan (2008) and Cao et al. (2009), see also the discussion on the projection estimator in Section 4.3.4. For working models $\pi(\boldsymbol{X};\boldsymbol{\psi})$ and $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$, this would proceed by assuming that $\pi(\boldsymbol{X};\boldsymbol{\psi})$ is correctly specified and we would estimate $\boldsymbol{\psi}$ by means of the MLE $\hat{\boldsymbol{\psi}}_n^{\text{MLE}}$. The parameter $\boldsymbol{\tau}$ would then be estimated via an estimator $\hat{\boldsymbol{\tau}}_n^{\text{EEM}}$, converging to a value $\boldsymbol{\tau}_{\text{EEM}}^*$ that minimizes the asymptotic variance of the doubly robust estimator using the CPI working model $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$ (assuming

$\pi(\boldsymbol{X};\boldsymbol{\psi})$ is correctly specified), even under misspecification of that CPI working model, but converging to $\boldsymbol{\tau}_0$ under correct specification of $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$. To focus ideas, suppose that the propensity score working model is fully specified by the function $\pi(\boldsymbol{X})$, so that $\pi(\boldsymbol{X}) = \pi_0(\boldsymbol{X})$ under correct specification. In this case, the asymptotic variance of the doubly robust estimator (using an estimator $\hat{\boldsymbol{\tau}}_n$ with probability limit $\boldsymbol{\tau}^*$) is given by the variance of its influence function

$$\begin{aligned}\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0) &= \frac{1-A}{1-\pi_0(\boldsymbol{X})}a_1^{(0)}(Y) - \nu_0 + \frac{A}{\pi_0(\boldsymbol{X})}a_2^{(0)}(Y) - \nu_0 \\ &\quad +\{A-\pi_0(\boldsymbol{X})\}\tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X}) \\ &= \frac{1-A}{1-\pi_0(\boldsymbol{X})}a_1^{(0)}\{Y(0)\} - \nu_0 + \frac{A}{\pi_0(\boldsymbol{X})}a_2^{(0)}\{Y(1)\} - \nu_0 \\ &\quad +\{A-\pi_0(\boldsymbol{X})\}\tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X}),\end{aligned}$$

where we used that $Y = Y(a)$ iff $A = a$ (consistency assumption) and with $a_1^{(0)}(Y) = E\{A^*I(Y \preceq Y^*)/\pi_0(\boldsymbol{X}^*)|Y\}$, $a_2^{(0)}(Y) = E[(1-A^*)I(Y^* \preceq Y)/\{1-\pi_0(\boldsymbol{X}^*)\}|Y]$ and

$$\tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X}) = \left[\frac{E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})} - \frac{E\{m(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*)|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\right].$$

This variance can be easily calculated using the law of iterated variance,

$$\begin{aligned}&\mathrm{var}\left\{\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\right\} \\ &\quad= \mathrm{var}\left[E\left\{\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\middle|\boldsymbol{X},Y(0),Y(1)\right\}\right] \\ &\qquad + E\left[\mathrm{var}\left\{\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\middle|\boldsymbol{X},Y(0),Y(1)\right\}\right] \\ &\quad= \mathrm{var}\left[a_1^{(0)}\{Y(0)\} + a_2^{(0)}\{Y(1)\}\right] \\ &\qquad + E\left(\pi_0(\boldsymbol{X})\{1-\pi_0(\boldsymbol{X})\}\left[\frac{a_1^{(0)}\{Y(0)\}}{1-\pi_0(\boldsymbol{X})} - \frac{a_2^{(0)}\{Y(1)\}}{\pi_0(\boldsymbol{X})} - \tilde{\alpha}_{\boldsymbol{\tau}^*}(\boldsymbol{X})\right]^2\right).\end{aligned}$$

From this, it follows that the value $\boldsymbol{\tau}^*_{\mathrm{EEM}}$ that minimizes this variance can by found as the solution to the equation $\partial\mathrm{var}\{\phi^{(0)}_{\tilde{\alpha}_{\boldsymbol{\tau}^*}}(Y,A,\boldsymbol{X};\nu_0)\}/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\boldsymbol{\tau}^*_{\mathrm{EEM}}} = \mathbf{0}$, which is

proportional to

$$E\Bigg(\pi_0(\boldsymbol{X})\{1-\pi_0(\boldsymbol{X})\} \times\left[\frac{a_1^{(0)}\{Y(0)\}}{1-\pi_0(\boldsymbol{X})}-\frac{a_2^{(0)}\{Y(1)\}}{\pi_0(\boldsymbol{X})}-\tilde{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})\right]\left.\frac{\partial\tilde{\alpha}_{\boldsymbol{\tau}}(\boldsymbol{X})}{\partial\boldsymbol{\tau}}\right|_{\boldsymbol{\tau}=\boldsymbol{\tau}^*_{\text{EEM}}}\Bigg)=\mathbf{0}.$$

We thus need to identify suitable estimating equations to obtain an estimator $\hat{\boldsymbol{\tau}}_n^{\text{EEM}}$ for the parameter indexing the PIM $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$ so that $\hat{\boldsymbol{\tau}}_n^{\text{EEM}}$ converges in probability to $\boldsymbol{\tau}^*_{\text{EEM}}$ under PIM misspecification and converges to $\boldsymbol{\tau}_0$ under a correctly specified PIM (but potentially misspecified propensity score working model). For this purpose, note that the aforementioned equation can be written as

$$\begin{aligned}&E\left(\pi_0(\boldsymbol{X})\left[a_1^{(0)}\{Y(0)\}-E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})|\boldsymbol{X}\}\right]\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})\right)\\&-E\left(\{1-\pi_0(\boldsymbol{X}^*)\}\left[a_2^{(0)}\{Y^*(1)\}-E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})|\boldsymbol{X}^*\}\right]\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X}^*)\right),\end{aligned}$$

with

$$\begin{aligned}\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})&=\left.\frac{\partial\tilde{\alpha}_{\boldsymbol{\tau}}(\boldsymbol{X})}{\partial\boldsymbol{\tau}}\right|_{\boldsymbol{\tau}=\boldsymbol{\tau}^*_{\text{EEM}}}\\&=\frac{E\{m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})|\boldsymbol{X}\}}{1-\pi_0(\boldsymbol{X})}-\frac{E\{m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X}^*,\boldsymbol{X};\boldsymbol{\tau}^*_{\text{EEM}})|\boldsymbol{X}\}}{\pi_0(\boldsymbol{X})}\end{aligned}$$

and $m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})=\partial m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})/\partial\boldsymbol{\tau}|_{\boldsymbol{\tau}=\boldsymbol{\tau}^*_{\text{EEM}}}$. Next, observe that

$$\begin{aligned}&E\left[\pi_0(\boldsymbol{X})a_1^{(0)}\{Y(0)\}\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})\right]\\&\quad=E\left[\pi_0(\boldsymbol{X})\frac{(1-A)}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}I(Y\preceq Y^*)\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})\right],\\&E\left[\{1-\pi_0(\boldsymbol{X}^*)\}a_2^{(0)}\{Y^*(1)\}\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X}^*)\right]\\&\quad=E\left[\{1-\pi_0(\boldsymbol{X}^*)\}\frac{(1-A)}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}I(Y\preceq Y^*)\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X}^*)\right],\\&E\left[\pi_0(\boldsymbol{X})E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})|\boldsymbol{X}\}\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})\right]\end{aligned}$$

$$= E\left[\pi_0(\boldsymbol{X})\frac{(1-A)}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})\right],$$

$$E\left[\{1-\pi_0(\boldsymbol{X}^*)\}E\{m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})|\boldsymbol{X}^*\}\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X}^*)\right]$$

$$= E\left[\{1-\pi_0(\boldsymbol{X}^*)\}\frac{(1-A)}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X}^*)\right].$$

From these calculations, it now follows that the value $\boldsymbol{\tau}^*_{\text{EEM}}$ must solve the equation

$$E\left(\frac{(1-A)}{1-\pi_0(\boldsymbol{X})}\frac{A^*}{\pi_0(\boldsymbol{X}^*)}\{I(Y\preceq Y^*)-m(0,1,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau}^*_{\text{EEM}})\}\right.$$
$$\left.\times\left[\pi_0(\boldsymbol{X})\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X})-\{1-\pi_0(\boldsymbol{X}^*)\}\overline{\alpha}_{\boldsymbol{\tau}^*_{\text{EEM}}}(\boldsymbol{X}^*)\right]\right)=\mathbf{0}.$$

We therefore propose to estimate $\boldsymbol{\tau}$ by means of the estimator $\hat{\boldsymbol{\tau}}^{\text{EEM}}_n$, which is defined as the solution to the estimating equation

$$\sum_{i=1}^{n}\sum_{j\neq i}\left(\frac{1-A_i}{1-\pi(\boldsymbol{X}_i)}\frac{A_j}{\pi(\boldsymbol{X}_j)}\{I_{ij}-m(0,1,\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\boldsymbol{\tau}}^{\text{EEM}}_n)\}\right.$$
$$\left.\times\left[\pi(\boldsymbol{X}_i)\widehat{\overline{\alpha}}_{\hat{\boldsymbol{\tau}}^{\text{EEM}}_n}(\boldsymbol{X}_i)-\{1-\pi(\boldsymbol{X}_j)\}\widehat{\overline{\alpha}}_{\hat{\boldsymbol{\tau}}^{\text{EEM}}_n}(\boldsymbol{X}_j)\right]\right)=\mathbf{0},$$

with $I_{ij}=I(Y_i\preceq Y_j)$ and

$$\widehat{\overline{\alpha}}_{\hat{\boldsymbol{\tau}}^{\text{EEM}}_n}(\boldsymbol{X}_i)=(n-1)^{-1}\sum_{k\neq i}\left\{\frac{m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X}_i,\boldsymbol{X}_k;\hat{\boldsymbol{\tau}}^{\text{EEM}}_n)}{1-\pi(\boldsymbol{X}_i)}-\frac{m_{\boldsymbol{\tau}}(0,1,\boldsymbol{X}_k,\boldsymbol{X}_i;\hat{\boldsymbol{\tau}}^{\text{EEM}}_n)}{\pi(\boldsymbol{X}_i)}\right\}.$$

It follows that when the PIM $m(A,A^*,\boldsymbol{X},\boldsymbol{X}^*;\boldsymbol{\tau})$ is correctly specified (but the propensity score is potentially misspecified, so that $\pi(\boldsymbol{X})\neq\pi_0(\boldsymbol{X})$), $\hat{\boldsymbol{\tau}}^{\text{EEM}}_n\xrightarrow{p}\boldsymbol{\tau}_0$. Furthermore, when the PIM is misspecified but the propensity score is correctly specified, $\hat{\boldsymbol{\tau}}^{\text{EEM}}_n\xrightarrow{p}\boldsymbol{\tau}^*_{\text{EEM}}$. It remains to investigate however how this estimating equation can be efficiently solved numerically and how to adapt the procedure when a parametric working model $\pi(\boldsymbol{X};\boldsymbol{\psi})$ for the propensity score is posited (with $\boldsymbol{\psi}$ estimated via MLE $\hat{\boldsymbol{\psi}}^{\text{MLE}}_n$). The resulting doubly robust estimator would then be given by $\hat{v}^{\text{EEM}}_{n,\text{DR}}\equiv\hat{v}_{n,\text{DR}}(\hat{\boldsymbol{\psi}}^{\text{MLE}}_n,\hat{\boldsymbol{\tau}}^{\text{EEM}}_n)$.

The aforementioned alternatives are theoretically appealing by their defining properties. However, it remains to be seen how to best implement the resulting estimating equations. In future research, we thus plan to further develop these alternatives and evaluate the relative performance of these different proposals $\hat{v}^{\mathrm{T}}_{n,\mathrm{DR}}$, $\hat{v}^{\mathrm{WR}}_{n,\mathrm{DR}}$, $\hat{v}^{\mathrm{BR}}_{n,\mathrm{DR}}$ and $\hat{v}^{\mathrm{EEM}}_{n,\mathrm{DR}}$.

CHAPTER 8

Conclusion and Future Research

8.1 Conclusion

Estimators that enjoy double robustness, a property originally introduced in Scharfstein et al. (1999a), consistently estimate the parameter of interest when at least one of two working models is correctly specified, regardless of which. This makes the doubly robust estimator a potential compromise estimator amidst competing estimators that each rely on a single, but different working model. See Chapter 3 for an overview. The appeal of these doubly robust estimators surpasses the defining property of double protection against model misspecification; many doubly robust estimators are locally efficient within a broad class of estimators. Indeed, for instance in Chapter 2 and Chapter 3, we demonstrated the local efficiency of the doubly robust estimator of a population mean outcome in the presence of missing data, explainable by measured covariates. In Chapter 7, we constructed a doubly robust estimator of the MPI (7.1), which guarantees consistency under correct specification of a working model for the propensity score or for the conditional probabilistic index (CPI). This enables adjustment for confounding and is particularly useful to assess a causal effect in observational studies with highly skewed

outcome data, where an additive causal effect may not constitute a good effect size. Interestingly, this doubly robust estimator is also locally efficient under a semiparametric model that assumes a correctly specified propensity score working model, which we showed in Chapter 7. Efficiency is attained locally at correct specification of the CPI working model. Because of this, the use of doubly robust estimators has also been advocated in randomized trial analyses: by exploiting the known randomization probabilities (and thus known propensity score), doubly robust estimators make it possible to increase power via covariate adjustment, without risking bias due to model misspecification; consistency is guaranteed by the known randomization probabilities. In Chapter 6 (see also Vermeulen et al. (2015)), we showed how we can implement such covariate adjustment to increase the power of the Mann-Whitney U test in randomized experiments.

Despite the attractive properties of doubly robust estimators, they have been the subject of recent debate (Kang and Schafer 2007a). As shown in Chapter 3, the asymptotic behavior of the doubly robust estimator does depend on the choice of nuisance parameter estimators under misspecification of at least one working model, in contrast to the case where both working models are correctly specified. This implies that more subtle choices (than for instance standard MLE or least squares) can be made for nuisance parameter estimators. For an overview of such alternatives, we refer to the discussion of Chapter 3. Furthermore, it is likely that model misspecification affects all working models in practice, and thus the very premise that at least one of both working models is correctly specified, lives on shaky grounds, thereby potentially making the double-protection property of a more academic interest. Moreover, the performance of doubly robust estimator can sometimes be worse than that of competing estimators that do not enjoy the double protection property. This fact served as the main motivation for writing the central part of thesis.

In Chapter 4, we therefore investigated the usefulness of doubly robust estimators from the perspective that all models are wrong. We found that, surprisingly, some doubly robust estimators partially retain their robustness properties even under misspecification of both working models and we referred to these estimators as **bias-reduced doubly robust estimators**, originally introduced in Vermeulen and Vansteelandt (2015a). The bias-reduced doubly robust estimation strategy is a

simple and fairly generic estimation principle for the (finite-dimensional) nuisance parameters indexing each of the working models, with the defining property of locally minimizing the squared first-order asymptotic bias of the doubly robust estimator in the direction of the nuisance parameters. The bias-reduced doubly robust estimation procedure moreover delivers estimators with a first-order ancillarity property with respect to the nuisance parameters and thus entails a simple asymptotic distribution of the doubly robust estimator. It is illustrated in Chapter 4 and 5 through extensive simulation studies that this novel estimation principle can result in drastic bias reductions and efficiency improvements as compared to alternative doubly robust estimators. Besides estimation of a population mean outcome susceptible to missingness, explainable by auxiliary covariates, we also illustrated the bias-reduced doubly robust estimation principle to the estimation of marginal treatment effects, G-estimation for semiparametric regression models and estimation of a population mean outcome, but when missingness is non-ignorable. We moreover extended this principle to multiply robust estimation in semiparametric interaction models. In Chapter 5, we additionally investigated how we can extend the original bias-reduced doubly robust estimation principle, which is restricted to the use of parametric nuisance working models, to the use of data-adaptive learning algorithms, in an attempt to allow for further bias reduction.

8.2 Future Research

In this section, we posit several remaining open questions and briefly make suggestions how we could solve these in future research.

8.2.1 Bias-reduced doubly robust estimation

Although the bias-reduced doubly robust estimation principle applies quite generically to many doubly robust estimators, it also comes with some limitations. Specifically, the procedure

(1) is currently restricted to one-dimensional target parameters;

(2) demands working models with nuisance parameters of the same dimension;

(3) is restricted to functionals with influence function that is linear in the target parameter.

In future research, we aim to develop extensions of the bias-reduced doubly robust estimation principle to overcome these limitations. Below, we briefly elaborate on possible suggestions to establish these extensions, but first, we motivate the practical importance of solutions to these limitations.

Application to longitudinal marginal structural models

The solutions to these problems will be important to enhance the performance of doubly robust estimators of the parameters indexing longitudinal Marginal Structural Models (MSMs) (Yu and van der Laan 2005). The parameters indexing such MSMs are typically multi-dimensional (see for instance Section 4.5.1 for an example of a linear MSM in a point-treatment study, where the parameter of interest is in fact two-dimensional). These marginal structural models have become very popular in the epidemiological literature to adjust for time-varying confounding in longitudinal observational studies. However, despite their popularity, these estimators can suffer heavily from large finite-sample bias and imprecision as a result of highly variable weights. This is especially common in studies where the exposure is continuous, or where the exposure is strongly correlated with subject characteristics. Moreover, because the weights are accumulated over time, this instability gets reinforced in studies with long follow-up. Possible lack of performance is often concealed by heuristic data analysis procedures, such as artificial truncation of extreme weights. As we have seen in Chapter 4 and Chapter 5, the bias-reduced estimation principle could overcome these difficulties. However, not only because the MSM-parameters are typically multi-dimensional, but also because non-linear MSMs lead to influence functions that are non-linear in the target parameters, and because the nuisance working models needed to obtain the doubly robust estimators are typically of different dimensions, the bias-reduced doubly robust estimation principle is not readily applicable. The aforementioned extensions to the theory of bias-reduced doubly robust estimation are thus necessary to handle the complexities of such longitudinal MSMs to enhance doubly robust estimation of these MSM-parameters.

Suggestions to establish the needed extensions

We now briefly describe some suggestions to overcome the aforementioned limitations of the bias-reduced doubly robust estimation procedure.

If interest lies in a multi-dimensional target parameter, the first problem (1) can be tackled by applying the bias-reduced estimation procedure for each target parameter separately, based on its influence function considering the other target parameters to be unknown. This may imply that different nuisance parameter estimators are used to estimate the different target parameters to optimize the performance of each parameter of interest in terms of bias (for a preliminary example, see Section 4.5.1). However, it may pose difficulties with respect to joint inference for all target parameters, which should be investigated in future research.

As argued in Section 4.7, in simple cases, the second problem (2) can be remedied by enlarging the working models until they reach the same dimension, which can be done using cleverly chosen covariates whose inclusion in the working model improves the performance of the doubly robust estimator for the target parameter according to a certain criterion. Unfortunately, this is somewhat ad-hoc and becomes prohibiting in more complex settings, such as longitudinal marginal structural models. One strategy to overcome this, is to locally minimize the bias of the doubly robust estimator in the direction of the nuisance parameters indexing only one working model rather than both. In Chapter 5, see also Vermeulen and Vansteelandt (2015b), some work related to this problem has already been done. In that chapter, the bias-reduced estimation principle is extended to incorporate the use of data-adaptive estimators by applying bias-reduction only in the direction of the parameters describing the finite-dimensional working model for the propensity score and desirable finite-sample performance of these estimators is seen in the simulation studies of Chapter 5. However, when applying bias-reduction in only one direction, open problems are that the asymptotic properties of the resulting estimators are less well understood, and moreover, that the resulting procedure is not unique. Indeed, it differs depending on whether or not one explicitly acknowledges that the influence function of the target parameter may depend on one of the nuisance parameters through its functional relation with the estimator of the other nuisance parameters. A further problem is how to best estimate the nuisance parameter indexing the

other working model. Alternatively, when bias reduction is considered in the direction of both nuisance parameters, it delivers too many estimating function for one nuisance parameter and too few for the other. This could be accommodated by minimizing a distance function based on the estimating functions for the first nuisance parameter, and supplementing additional estimating functions for the other. Open problems are how to best choose this distance function as well as additional estimating functions. Solutions to these problems are important, not only because it will allow for nuisance working models of unequal dimensions, but also to allow for infinite-dimensional nuisance working models that are fitted via machine learning algorithms, as already suggested in Chapter 5.

Because the bias-reduced estimation principle applied to a functional with influence function non-linear in the target parameter requires knowledge of the unknown value of the target parameter (see for instance the example of G-estimation in semiparametric log-linear models on page 104), its current use is prohibited to functionals with influence function linear in the target parameter, problem (3). In future research, we will try to overcome this problem by applying the bias-reduced estimation principle for each fixed choice of the target parameter over some grid, to minimize the bias of the expectation of the influence function itself under double misspecification. We will examine the asymptotic properties of the doubly robust estimator obtained by inverting that expectation, with the nuisance parameters substituted by the estimators defining the bias-reduced estimating principle.

Our aim is to solve these remaining open problems in future research.

8.2.2 Extensions to the Mann-Whitney U test

In this section, we briefly discuss possible solutions to several remaining open questions concerning the proposed extensions to the Mann-Whitney U test.

The extended Mann-Whitney U test in randomized experiments

In the discussion of Chapter 6, we noted that a limitation of the proposed permutation test, that extends the classical Mann-Whitney U test by allowing for covariate adjustment, may lead to a severe decrease in sample size when there is substantial missingness in the covariates and a complete-case analysis is performed. An

efficiency benefit as compared to an unadjusted analysis may then be lost. We indicated that in such case, we recommend to use multiple imputation to deal with the missing covariate data. However, these imputations should be based on an imputation model that only includes covariates but no outcome or exposure. In this manner, we do not induce bias in the estimator of the MPI, even when the imputation model is misspecified and regardless of the missing data mechanism.

A next question would then be how to combine the results. Suppose that after the imputation procedure, we have M, e.g., $M = 5$, imputed datasets. Following Rubin's rule (Rubin 1987), the pooled (P) observed test statistic $\widehat{T}_{n,0}^{(\mathrm{P})}$ can be easily obtained by averaging the observed test statistics $\widehat{T}_{n,0}^{(j)}$ ($j = 1, \ldots, M$), obtained for each of the imputed datasets:

$$\widehat{T}_{n,0}^{(\mathrm{P})} = M^{-1} \sum_{j=1}^{M} \widehat{T}_{n,0}^{(j)}.$$

Calculating a *pooled p-value* is less straightforward, because each of the p-values obtained from the imputed datasets are underestimated. A combined p-value could however be obtained by stacking the M permutation null distributions obtained from each of the M different imputed datasets and regarding these stacked permutation null distributions as the permutation null distribution. In this manner, we both acknowledge the within and between imputation variance of the observed test statistic. In future research, we plan to investigate the practical efficiency benefit from this procedure as compared to an unadjusted analysis.

The extended Mann-Whitney U test in observational studies

Another open question arises from Chapter 7. In the discussion of this chapter, we constructed several alternative nuisance parameter estimators of the nuisance parameters indexing the working models used in the definition of the doubly robust estimator of the MPI, extending the classical Mann-Whitney U test, by enabling adjustment for confounding in observational studies. These alternatives are theoretically appealing. However, it remains to be seen how to best implement the resulting estimating equations. Furthermore, empirical validation is also needed to compare the relative performance of these alternative doubly robust estimators

$\hat{\nu}^{\mathrm{WR}}_{n,\mathrm{DR}}$, $\hat{\nu}^{\mathrm{BR}}_{n,\mathrm{DR}}$ and $\hat{\nu}^{\mathrm{EEM}}_{n,\mathrm{DR}}$ to the performance of the *standard* doubly robust estimator $\hat{\nu}^{\mathrm{T}}_{n,\mathrm{DR}}$ and to the performance of the simpler estimators $\hat{\nu}_{n,\mathrm{IMP}}$ and $\hat{\nu}_{n,\mathrm{IPTW}}$.

A further research question would be to investigate how we could extend the counterfactual definition of the MPI (7.1) to marginal structural models for certain conditional probabilistic indices and how to obtain doubly robust estimators for these MSM-parameters.

CHAPTER 9

Samenvatting

Het schatten van heel wat statistische parameters vereist het postuleren van zogenoemde **nuisance working modellen**. Deze modellen zijn niet van wetenschappelijk belang. Ze zijn echter nodig om een schatter te bekomen voor de parameter waarin we geïnteresseerd zijn, de **doel-parameter**, zodanig dat deze stabiele eigenschappen heeft in kleine tot middelgrote steekproeven. Dit wordt ook wel de **curse of dimensionality** genoemd (Robins and Ritov 1997) omdat niet-parametrische modellen niet mogelijk zijn met hoger-dimensionale covariaten. Bijvoorbeeld, beschouw een studie waarbij we geïnteresseerd zijn in het schatten van het populatie gemiddelde van een bepaalde uitkomst Y, maar waarbij we deze uitkomst niet voor elk individu in onze steekproef observeren. Veronderstel echter dat we voor elk individu wel een set van covariaten $\boldsymbol{X}$ hebben verzameld met de eigenschap dat, gegeven deze covariaten (dus binnen een subgroep van individuen met hetzelfde covariaten-patroon), het al dan niet ontbrekend zijn van de uitkomst, als volledig willekeurig kan gezien worden. Het correct inschatten van het populatie gemiddelde van de uitkomst Y vergt dan dat we ofwel de relatie modelleren tussen de uitkomst en de covariaten (via een uitkomst working model), ofwel de relatie modelleren tussen het al dan niet ontbrekend zijn van de uitkomst en de covariaten (via een

missingness working model). Een ander voorbeeld is een typisch probleem uit de causale besluitvorming: het schatten van een causaal effect van een bepaalde blootstelling A op een uitkomst Y, waarbij we een rijke set van confounders $\boldsymbol{X}$ tot onze beschikking hebben. Hierbij zijn confounders variabelen die zowel de blootstelling als de uitkomst beïnvloeden. Om het causaal effect van A op Y correct te kunnen inschatten, moeten we ofwel een working model voor de relatie tussen de uitkomst Y en de set van confounders $\boldsymbol{X}$ en de behandeling A postuleren (het uitkomst working model), ofwel moeten we een working model voor de blootstelling A gegeven de confounders $\boldsymbol{X}$ (het propensity score working model) postuleren. Een oprechte bezorgdheid is echter dat misspecificatie van deze nuisance working modellen bias kan introduceren in de schatter van de parameter waarin we geïnteresseerd zijn (Robins 1999a). Deze vorm van bias wordt **model-misspecificatie bias** genoemd.

In heel wat schattingsproblemen met ontbrekende gegevens en schattingsproblemen in de causale besluitvorming kan deze zorg voor model-misspecificatie bias verlicht worden via het gebruik van zogenoemde **dubbel robuuste schatters**. Deze schatters verzwakken de afhankelijkheid van model assumpties omdat ze de mogelijkheid bieden om niet beperkt te zijn tot één bepaald nuisance working model. Ze vergen daarentegen specificatie van ten minste twee nuisance working modellen, waarvan slechts één correct gespecificeerd moet zijn om een consistente schatter voor de doel-parameter te bekomen (Scharfstein et al. 1999a; Robins and Rotnitzky 2001). Bijvoorbeeld, een dubbel robuuste schatter voor een causaal effect vergt zowel een uitkomst working model als een propensity score working model en is consistent zodra ten minste één van deze twee working modellen correct gespecificeerd is. Dergelijke schatters geven de data-analist dus twee kansen om een correcte inschatting te bekomen voor de doel-parameter. Dubbel robuuste schatters kunnen bijgevolg gezien worden als een compromis tussen schatters die gebaseerd zijn op slechts één working model. Bovendien maken dubbel robuuste schatters in heel wat gevallen optimaal gebruik van de beschikbare informatie in de data in de zin dat ze **lokaal efficiënt** zijn binnen een grote klasse van schatters (Bickel et al. 1993a). Bijvoorbeeld, een dubbel robuuste schatter voor een causaal effect heeft de kleinste variantie van alle schatters die consistent zijn onder een correct gespecificeerd propensity score working model, gegeven dat ook het uitkomst working model correct gespecificeerd is; efficiëntie wordt

dus lokaal bereikt. Daarom wordt het gebruik van dubbel robuuste schatters ook aangemoedigd in de analyse van gerandomiseerde studies waarbij de propensity score, of dus de randomisatie kans, gekend is (Tsiatis et al. 2008; Moore and van der Laan 2009; Vermeulen et al. 2015).

Het schatten van de nuisance parameters (de parameters die de nuisance working modellen beschrijven) kreeg in het verleden echter relatief weinig aandacht. Dit omdat theoretische resultaten aantonen dat de keuze van schatters voor deze nuisance parameters (zolang deze wortel-*n* consistent zijn) geen invloed heeft op de asymptotische variantie van de dubbel robuuste schatter wanneer beide nuisance working modellen correct gespecificeerd zijn. Dit leidde ondermeer tot het default gebruik van maximum kans schatters voor de nuisance parameters (Bang and Robins 2005).

Recentelijk werden dubbel robuuste schatters echter een belangrijk onderwerp van discussie (Kang and Schafer 2007a; Ridgeway and McCaffrey 2007; Robins et al. 2007; Tan 2007; Tsiatis and Davidian 2007; Kang and Schafer 2007b). Eerst en vooral, wanneer ten minste één working model niet correct gespecificeerd is, dan kunnen in principe voor een gegeven doel-parameter oneindig veel dubbel robuuste schatters geconstrueerd worden. Dit is mogelijk door de keuze van schatters voor de nuisance parameters te laten variëren. Hierbij heeft elke geconstrueerde dubbel robuuste schatter mogelijks een heel ander gedrag onder working model misspecificatie. Dit impliceert dat meer subtiele keuzes voor schatters van de nuisance parameters kunnen worden gemaakt. Hierbij worden deze schatters zodanig geconstrueerd dat ze de performantie van de dubbel robuuste schatter verbeteren onder working model misspecificatie. Ten tweede is het heel waarschijnlijk dat in de praktijk beide nuisance working modellen niet correct gespecificeerd zijn. Dit betekent dat de premisse dat één van beide modellen correct gespecificeerd zou zijn op losse schroeven komt te staan. Bovendien kan onder dubbele working model misspecificatie, de performantie van een dubbel robuuste schatter slechter zijn dan de performantie van een meer simpele schatter die gebaseerd is op slechts één working model en dus niet van deze dubbele bescherming geniet.

Deze problemen motiveerden echter heel wat statistici om alternatieve schatters voor de nuisance parameters te identificeren, die voornamelijk variantie-reductie onder misspecificatie van één working model beogen maar ook hoe slim gebruik

kan gemaakt worden van data-adaptieve schattingstechnieken. In deze thesis onderzoeken we echter in eerste instantie het nut van dubbel robuuste schatters vanuit het perspectief dat beide nuisance working modellen fout zijn. Het voornaamste doel van deze thesis is dus om een algemene schattings-strategie te ontwikkelen voor de nuisance working modellen die nodig zijn in de constructie van een dubbel robuuste schatter, die **bias-reductie** beoogt, vanuit het perspectief dat beide working modellen fout zijn. Dit is gemotiveerd door het feit dat de bias van een dubbel robuuste schatter voornamelijk groot kan worden onder misspecificatie van beide nuisance working modellen.

Omdat heel wat van de resultaten die in deze thesis worden ontwikkeld, gebaseerd zijn op de theorie van semi-parametrische modellen en semi-parametrische efficiëntie, frissen we beknopt enkele basisprincipes van de **semi-parametrische theorie** op in **Hoofdstuk 2**. In Sectie 2.1 starten we met het herhalen van de definitie van een statistisch model. Meer specifiek zetten we het verschil tussen niet-parametrische, semi-parametrische en parametrische modellen in de verf. Omdat we de semi-parametrische theorie zullen bekijken vanuit een geometrisch perspectief, vatten we de relevante theorie omtrent **Hilbertruimtes** samen in Sectie 2.2. Vervolgens bouwen we verder op deze geometrische concepten en introduceren we in Sectie 2.3 de theorie van **invloedsfuncties** in parametrische (en dus eindig-dimensionale) modellen. Dit laat toe om de efficiëntie van reguliere en asymptotisch normale schatters te bestuderen en om zo de **efficiëne invloedsfunctie** te identificeren (de invloedsfunctie met de kleinste variantie). In Sectie 2.4 breiden we deze ideeën dan uit tot semi-parametrische (en dus oneindig-dimensionale) modellen. We eindigen dit hoofdstuk in Sectie 2.5 met het toepassen van de semi-parametrische theorie op het probleem van het schatten van een populatie gemiddelde van een uitkomst waarbij we deze uitkomst niet voor elk individu uit onze steekproef observeren maar waarbij we wel een set van covariaten ter beschikking hebben die het ontbrekend zijn van de uitkomst kunnen verklaren. In het bijzonder identificeren we de efficiënte invloedsfunctie en een corresponderende lokaal efficiënte reguliere en asymptotisch normale schatter van deze doel-parameter. Deze is gebaseerd op een working model voor het missingness mechanisme (de kans op een ontbrekende uitkomst, gegeven de covariaten) en een working model voor het conditioneel gemiddelde van de uitkomst, gegeven de covariaten. Deze schatter zal doorheen

deze thesis gebruikt worden als studieobject en zijn eigenschappen zullen bovendien in detail worden besproken. De inhoud van dit hoofdstuk is voornamelijk gebaseerd op het uitstekende boek *Semiparametric Theory and Missing Data* door Tsiatis A.A. (2006), Springer: New York. Dit boek geeft een meer gedetailleerde beschrijving van de semi-parametrische efficiëntie theorie in problemen met ontbrekende gegevens.

In **Hoofdstuk 3** demonstreren we dat de lokaal efficiënte schatter (geconstrueerd in Hoofdstuk 2, Sectie 2.5) ook een andere merkwaardige eigenschap bezit: **dubbel robuustheid**. Deze eigenschap betekent dat de schatter consistent is voor de doel-parameter wanneer ofwel het working model voor het missingness mechanisme, ofwel het working model voor het conditioneel gemiddelde van de uitkomst, correct gespecificeerd is. Dit wordt expliciet aangetoond in Sectie 3.3. Dubbel robuustheid is niet beperkt tot dit voorbeeld. Tegenwoordig zijn er heel wat dubbel robuuste schatters geconstrueerd voor een variëteit aan statistische parameters. In Sectie 3.3 geven we een overzicht van bestaande dubbel robuuste schatter in de literatuur van de ontbrekende gegevens en causale besluitvorming. Hun populariteit kan worden gestaafd aan de hand van heel wat wetenschappelijke artikels die dubbel robuuste schatters behandelen: meer dan 2000 op Google Scholar en meer dan 200 op Web of Science, ondanks dat deze theorie omtrent dubbel robuustheid eigenlijk nog relatief nieuw is. Recentelijk worden dubbel robuuste schatters ook overwogen door grote bedrijven, zoals Google en Microsoft, in de context van beleids-optimalisatie en evaluatie van inhoudelijke aanbevelingen en reclame op het internet (Dudík et al. 2015). Ondanks de dubbele bescherming tegen model misspecificatie, waarschuwen Kang and Schafer (2007a) voor mogelijks rampzalige performantie van bepaalde dubbel robuuste schatter (relatief ten opzichte van meer simpele schatters) wanneer ten minste één van beide working modellen incorrect gespecificeerd zijn. Ze brengen bovendien aan het licht dat, voor een gegeven doel-parameter, heel wat verschillende dubbel robuuste schatters geconstrueerd kunnen worden, die elk mogelijks een heel ander gedrag en heel andere eigenschappen kunnen vertonen onder misspecificatie van tenminste één working model. Daarom bestuderen we in Sectie 3.4 in detail de asymptotische verdeling van dubbel robuuste schatters onder mogelijke misspecificatie van de (eindig-dimensionale) working modellen. Deze problematiek zorgde voor de ontwikkeling van heel wat

alternatieve schattingsmethoden voor de nuisance parameters die deze working modellen beschrijven. Deze focussen voornamelijk op variantie-reductie onder misspecificatie van één working model en hoe men slim gebruik kan maken van data-adaptieve schatters. In Sectie 3.5 geven we een overzicht van dergelijke bestaande alternatieven.

In deze thesis zullen we ons echter focussen op **bias-reductie**, eerder dan variantie-reductie. Meer specifiek, in **Hoofdstuk 4** bestuderen we het nut van dubbel robuuste schatters vanuit het perspectief dat beide working modellen incorrect gespecificeerd zijn. Dit wordt gestimuleerd door het feit dat de bias van een dubbel robuuste schatter voornamelijk hoog kan zijn onder dubbele working model misspecificatie. In het bijzonder stellen we in Sectie 4.2 een redelijk simpele en generieke alternatieve schattings-strategie voor de nuisance parameters van alle working modellen voor. Deze schattings-strategie heeft de definiërende eigenschap dat de kwadratische eerste orde asymptotische **bias** van de dubbel robuuste schatter onder dubbele working model misspecificatie **lokaal wordt geminimaliseerd**, in de richting van de nuisance parameters. Deze procedure wordt daarom **bias-gereduceerd dubbel robuust schatten** genoemd. Naast bias reductie bezit de bias-gereduceerde dubbel robuuste schatter ook een eerste orde ancillariteits-eigenschap met betrekking tot de nuisance parameters en levert dus een dubbel robuuste schatter met een eenvoudige asymptotische verdeling en makkelijk te berekenen standaard errors. In Sectie 4.3 passen we de bias-gereduceerde schattings-methode toe op het missing data probleem, geïntroduceerd in Hoofdstuk 2. In Sectie 4.4 illustreren we, door middel van uitgebreide simulatie studies, dat deze nieuwe schattings-strategie kan resulteren in aanzienlijke bias reducties en efficiëntie verbeteringen, in vergelijking met bestaande alternatieve dubbel robuuste schatters. In Sectie 4.5 bestuderen we enkele andere dubbel en meervoudig robuuste schatters, zoals het schatten van een marginaal behandelingseffect in observationele studies, G-estimation in semi-parametrische regressie modellen, het schatten van een populatie gemiddelde van een bepaalde uitkomst wanneer het ontbrekend zijn van de uitkomst niet kan verklaard worden louter op basis van een set van covariaten en tonen we bovendien aan hoe de bias-gereduceerde strategie kan uitgebreid worden tot meervoudig robuuste schatters in semi-parametrische interactie modellen. We besluiten dit hoofdstuk met een analyse van de SUPPORT-data in Sectie 4.6 en een discussie in

Sectie 4.7. In de Appendix van dit hoofdstuk voorzien we ook een R-functie om de bias-gereduceerde dubbel robuuste schatter te bekomen voor het beschouwde missing data probleem, bestudeerd in Sectie 4.3.

De bias-gereduceerde dubbel robuuste schattings-strategie die we introduceerden in Hoofdstuk 4, is jammergenoeg beperkt tot het gebruik van parametrische nuisance working modellen. In **Hoofdstuk 5** onderzoeken we hoe we **data-adaptieve schatters** voor de nuisance working modellen kunnen integreren in de bias-gereduceerde schattings-procedure om zo nog verdere bias-reductie toe te laten. De voorgestelde strategie, die we **data-adaptief bias-gereduceerd dubbel robuust schatten** noemen, leidt bovendien ook tot een oplossing voor één van de beperkingen van het originele voorstel. Het is namelijk niet meer nodig dat beide nuisance working modellen van een gelijke dimensie zijn. In Sectie 5.4 introduceren we deze strategie in detail voor het missing data probleem dat geïntroduceerd werd in Hoofdstuk 2. In Sectie 5.5 illustreren we de goede performantie van de voorgestelde procedure, in vergelijking met andere alternatieven. We eindigen dit hoofdstuk in Sectie 5.6 waarin we het vrij generieke karakter van de data-adaptieve bias-gereduceerde dubbel robuuste schattings-techniek illustreren door dit toe te passen op een lineaire instrumentele variabele analyse. In de Appendix van Hoofdstuk 5 voorzien we ook een R-functie om de data-adaptieve bias-gereduceerde dubbel robuuste schatter te bekomen voor het beschouwde missing data probleem uit Sectie 5.4.

In het eerste deel van deze thesis hebben we ons vooral gefocussed op het verbeteren van bestaande dubbel robuuste schatters onder nuisance working model misspecificatie. In het tweede deel van deze thesis zullen we ons echter focussen op de constructie van nieuwe dubbel robuuste schatters. Meer specifiek zulle we in **Hoofdstuk 6** de lokale efficiëntie eigenschap exploiteren, een eigenschap die vele dubbel robuuste schatters bezitten. Dit zullen we doen in de context van de **Mann-Whitney U-test**. Deze test wordt frequent gebruikt om behandelingseffecten te detecteren in de analyse van gerandomiseerde experimenten waarbij de uitkomst scheef verdeeld is of waarbij de steekproef klein is. In de praktijk wordt echter routinematig ook achtergrondinformatie verzameld door middel van baseline covariaten, zodat de power van de standaard Mann-Whitney U test nog verbeterd kan worden, vermits deze test dergelijke informatie negeert. Het exploiteren van deze

informatie wordt traditioneel gedaan door deze covariaten toe te voegen aan een regressiemodel, wat resulteert in conditionele effecten. In het bijzonder kan dit, in de context van de Mann-Whitney U test, gedaan worden via **probabilistische index modellen** (PIMs). Deze modelleren de kans dat de uitkomst van een willekeurig persoon die behandeld is hoger is dan de uitkomst van een ander willekeurig persoon die niet behandeld is. Hoe dit kan gedaan worden, demonstreren we in Sectie 6.3. Jammergenoeg kan deze procedure echter leiden tot bias in de schatting voor het behandelingseffect wanneer we deze PIM niet correct specificeren. Bovendien blijft het echter een subtiele vraag of het corrigeren voor baseline covariaten in niet-lineaire regressiemodellen wel degelijk leidt tot het verhogen van de power om een behandelingseffect te detecteren en bovendien, door de non-collapsibility van niet-lineaire effect-maten, verandert de interpretatie van het behandelingseffect wanneer we corrigeren voor verschillende subsets van covariaten. Daarom argumenteren we om te focussen op de **marginale probabilistische index** (MPI), de effect maat die ook beoogt wordt door de standaard Mann-Whitney U test. In Sectie 6.4 tonen we aan hoe we een schatter kunnen bekomen voor de MPI door middel van het standardiseren van PIM-predicties. Bovendien merken we op dat via bepaalde fittings strategieën van de PIM, de bekomen schatter robuust is tegen misspecificatie van de PIM. Vervolgens, in Sectie 6.5, presenteren we de formele semiparametrische theorie die het voorgaande resultaat ondersteunt en identificeren we een **lokaal efficiëne schatter** voor de MPI. Deze is consistent onder een groot statistisch model dat enkel veronderstelt dat de behandeling en de covariaten onafhankelijk zijn van elkaar, wat gegarandeerd is door de randomisatie. Deze lokaal efficiënte schatter laat nu ondermeer toe om te corrigeren voor de baseline covariaten zonder het risico te lopen om bias te creëren door misspecificatie van de PIM. We tonen verder aan hoe we correcte asymptotische besluitvorming kunnen doen. Omdat de Mann-Whitney U test echter vaak wordt gebruikt in kleine steekproeven, waar de asymptotische sandwich schatter voor de standaard error niet noodzakelijk een goede benadering is voor de echte variabiliteit van de schatter, vooral wanneer ook covariaat-selectie wordt gedaan, construeren we ook een **permutatie test** in Sectie 6.6. Deze is gebaseerd op de lokaal efficiënte schatter voor de MPI. Interessant om op te merken is dat deze permutatie test toelaat om covariaat-selectie te doen zonder hierbij het behoudt van de Type I fout in het gedrang te brengen.

We illustreren de performantie van de voorgestelde procedures in een analyse van de ACTG 175-data in Sectie 6.7 en via uitgebreide simulatie studies in Sectie 6.8. We eindigen dit hoofdstuk met een discussie in Sectie 6.9. In de Appendix van Hoofdstuk 6, voorzien we een R-functie voor het schatten van de MPI en om asymptotische besluitvorming te doen. Bovendien voorzien we ook een R-functie om de voorgestelde permutatie test uit te voeren.

In **Hoofdstuk 7** breiden we de resultaten van Hoofdstuk 6 uit naar observationele studies: we stellen in het bijzonder een **dubbel robuuste uitbreiding van de Mann-Whitney** U **test** voor die toelaat om te corrigeren voor confounding. In Sectie 7.2 geven we de definitie van de MPI, gebaseerd op counterfactuals, en in Sectie 7.3 tonen we aan dat deze causale parameter kan geïdentificeerd worden op basis van de geobserveerde data wanneer we veronderstellen dat er geen ongemeten confounders zijn voor de relatie tussen de blootstelling en de uitkomst. Vervolgens construeren we in Sectie 7.4 een dubbel robuuste schatter voor de MPI zodanig dat deze een consistente schatter levert voor de MPI wanneer ofwel een working model voor de propensity score correct gespecificeerd is, ofwel een working model voor de conditionele probabilistische index correct gespecificeerd is. We leiden de asymptotische verdeling af van deze schatter in Sectie 7.5 en in Sectie 7.6 tonen we aan dat deze dubbel robuuste schatter lokaal efficiënt is onder een semiparametrisch model dat veronderstelt dat het model voor de propensity score correct gespecificeerd is, waarbij efficiëntie lokaal wordt bereikt wanneer ook het model voor de conditionele probabilistische index correct gespecificeerd is. We eindigen dit hoofdstuk met een discussie in Sectie 7.7, waar we verschillende alternatieve schatters voorstellen voor de nuisance parameters die de working modellen indexeren. Meer specifiek leggen we kort uit hoe we een dubbel robuuste regressie imputatie schatter kunnen bekomen, hoe we een bias-gereduceerde dubbel robuuste schatter kunnen bekomen en hoe we een dubbel robuuste schatter aan de hand van empirische efficiëntie maximalisatie kunnen bekomen.

We eindigen deze thesis in **Hoofdstuk 8** met een reflectie op de bekomen resultaten en een finale conclusie. Bovendien maken we een aantal suggesties voor toekomstig onderzoek.

CHAPTER 10

Scientific Ouput

Publications

Published

- Vermeulen, K., Thas, O., and Vansteelandt, S. (2015). "Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment," *Statistics in Medicine*, 34, 1012-1030.
- Vermeulen, K., and Vansteelandt, S. (2015). "Bias-Reduced Doubly Robust Estimation," *Journal of the American Statistical Association*, in press.
- Verbiest, N., Vermeulen, K., and Teredesai, A. (2014), *Data Classification: Algorithms and Applications*. In Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. *Book chapter: Evaluation of Classification Methods.*

Submitted

- Vermeulen, K., and Vansteelandt, S. (2015). "Data-Adaptive Bias-Reduced Doubly Robust Estimation," *The International Journal of Biostatistics*, sub-

mitted.

Presentations at Conferences

The results of my research were presented at several national and international conferences as well as informal meetings.

Contributed talks

- Joint Statistical Meetings 2012, July 28-August 2, 2012, San Diego, USA.
  Talk entitled: **On estimation of nuisance working models in doubly robust inference.** (contributed papers: Missing Data Methods)

- 20th Annual Meeting of the Belgian Statistical Society, October 24-25, 2012, Luik, Belgium.
  Talk entitled: **On estimation of nuisance working models in doubly robust estimators.** (contributed session: Robustness)

- 21st Annual Meeting of the Belgian Statistical Society, October 9-11, 2013, Ghent, Belgium.
  Talk entitled: **A doubly robust adaptation of the Mann-Whitney test with application to randomized clinical trials.** (contributed papers: Confounding)

- UK Causal Inference Meeting, April 28-29, 2014, Cambridge, UK.
  Talk entitled: **Focused estimation of nuisance parameters in doubly robust inference.** (contributed session: Identification, Robustness and DAGs)

- Joint Statistical Meetings 2014, August 2-7, 2014, Boston, USA.
  Talk entitled: **Increasing the power of the Mann-Whitney test through flexible covariate adjustment in randomized experiments.** (contributed papers: Causal Inference and Dynamic Treatment Regimens)

- UK Causal Inference Meeting 2015, April 15-17, 2015, Bristol, UK.
  Talk entitled: **Bias-Reduced Doubly Robust Estimation.**

Invited talks

- Joint Statistical Meetings 2013, August 3-8, 2013, Montréal, Canada.
  Talk entitled: **Improving the finite-sample performance of doubly robust estimators through FOCUSED nuisance parameter estimation.** (invited papers: Toward Better Statistical Methods for Causal Inference)
- IBS Channel 2015, April 20-22, 2015, Nijmegen, The Netherlands.
  Talk entitled: **Bias-Reduced Doubly Robust Estimation.** (invited session: Confounder Modeling and Selection)
- Joint Statistical Meetings 2015, August 8-13, 2015, Seattle, USA.
  Talk entitled: **Bias-Reduced Doubly Robust Estimation.** (topic contributed session: Fresh Perspectives in Causal Inference, III)

Poster presentations

- International Biometric Conference 2014, July 6-11, 2014, Florence, Italy.
  Poster entitled: **Increasing the power of the Mann-Whitney test through flexible covariate adjustment in randomized trials.**

Informal talks

- Informal Causal Inference Meeting, May 10, 2012, Ghent.
  Talk entitled: **On estimation of nuisance working models in doubly robust estimators.**
- Informal Causal Inference Meeting, March 1, 2013, Ghent.
  Talk entitled: **Targeted Maximum Likelihood Estimation (TMLE) for point-treatment data.**
- TWIST-symposium, March 4, 2014, Ghent.
  Talk entitled: **Focused estimation of nuisance parameters in doubly robust inference.**

Bibliography

Acion, L., Peterson, J. J., Temple, S., and Arndt, S. (2006), "Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects," *Statistics in Medicine*, **25**, 591–602.

Andrews, D. W. K. (1994a), "Asymptotics for Semiparametric Econometrics Models Via Stochastic Equicontinuity," *Econometrica*, **62**, 43–72.

— (1994b), "Empirical Process Methods in Econometrics," *Handbook of Econometrics*, **4**, 2247–2294.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000), "Subgroup analysis and other (mis)uses of baseline data in clinical trials," *The Lancet*, **21**, 1064–1069.

Bang, H. and Robins, J. M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, **61**, 962–972.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993a), *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press, Baltimore.

— (1993b), *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press, Baltimore.

Boos, D. D. and Stefanski, L. A. (2013), *Essential Statistical Inference*, Springer Texts in Statistics.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification And Regression Trees*, Belmont, CA: Wadsworth International Group.

Brumback, L. C., Pepe, M. S., and Alonzo, T. A. (2006), "Using the ROC curve for gauging treatment effect in clinical trials," *Statistics in Medicine*, **25**, 575–590.

Cao, W. H., Tsiatis, A. A., and Davidian, M. (2009), "Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data," *Biometrika*, **96**, 723–734.

Carpenter, J. R. and Kenward, M. G. (2008), *Missing data in randomized controlled trials – a practical guide*, Birmingham: National Institute for Health Research, Publication RM03/JH17/MK.

Chen, H. Y. (2007), "A Semiparametric Odds Ratio Model for Measuring Association," *Biometrics*, **63**, 413–421.

Chen, S. X., Qin, J., and Tang, C. Y. (2013), "Mann-Whitney test with adjustments to pretreatment variables for missing values and observational study," *Journal of the Royal Statistical Society Series B - Statistical Methodology*, **75**, 81–102.

Chung, E. and Romano, P. J. (2011), "Asymptotically valid and exact permutation tests based on two-sample U-statistics," Tech. Rep. 2011-09, Department of Statistics, Stanford University.

— (2013), "Exact and asymptotically robust permutation tests," *The Annals of Statistics*, **41**, 484–507.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., William, J. F., Vidaillet, H., Broste, S., Bellamy, P., Joanne, L., and Knaus, W. A. (1996), "The Effectiveness of Right Heart Catherization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association*, **276**, 889–897.

Cox, D. R. (1980), "Local Ancillarity," *Biometrika*, **67**, 279–286.

D'Agostino, R. B., Campbell, M., and Greenhouse, J. (2006), "The Mann-Whitney statistic: continuous use and discovery," *Statistics in Medicine*, **25**, 541–542.

d'Agostino, R. C. (1998), "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group," *Statistics in Medicine*, **17**, 2265–2281.

Davidian, M., Tsiatis, A. A., and Leon, S. (2005), "Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with Discussion)," *Statistical Science*, **20**, 261–301.

De Neve, J. and Sabbe, N. (2013), *An R package for fitting probabilistic index models: the* `pim` *package*.

Deville, J. D. and Särndal, C. E. (1992), "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, **87**, 376–382.

Díaz, I. and Rosenblum, M. (2014), "Targeted Maximum Likelihood Estimation using Exponential Families," *ArXiv e-prints*.

Dudík, M., Dumitru, E., Langford, J., and Li, L. (2015), "Doubly Robust Policy Evaluation and Optimization," *Statistical Science*, **29**, 485–511.

Ernst, M. D. (2004), "Permutation methods: a basis for exact inference," *Statistical Science*, **19**, 676–685.

Gill, R. D. (1989), "Non- and Semi-parametric Maximum Likelihood Estimators and the bond Mises Method (Part 1)," *Scandinavian Journal of Statistics*, **16**, 97–128.

Goetgeluk, S., Vansteelandt, S., and Goetghebeur, E. (2008), "Estimation of Controlled Direct Effects," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **70**, 1049–1066.

Greenland, S., Robins, J. M., and Pearl, J. (1999), "Confounding and collapsibility in causal inference," *Statistical Science*, **14**, 29–46.

Grissom, R. (1994), "Probability of the superior outcome of one treatment over another," *Journal of Applied Psychology*, **79**, 314–316.

Grouin, J. M., Day, S., and Lewis, J. (2004), "Adjustment for baseline covariates: an introductary note," *Statistics in Medicine*, **23**, 697–699.

Gruber, S. and van der Laan, M. (2010), "A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome," *The International Journal of Biostatistics*, **6**.

— (2014), *Targeted Maximum Likelihood Estimation: the* `tmle` *package*.

Hahn, J. (1998), "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, **66**, 315–331.

Hájek, J. (1970), "A characterization of limiting distributions of regular estimates," *Zeitschrift Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **14**, 323–330.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., Merigan, T. C., Blaschke, T. F., Simpson, D., McLaren, C., Rooney, J., and Salgo, M. (1996), "A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter," *New England Journal of Medicine*, **335**, 1081–1089.

Hand, D. (1992), "On comparing two treatments," *The American Statistician*, **46**, 190–192.

Hanley, J. and McNeil, B. (1982), "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, **143**, 29–36.

Hauck, W. W., Anderson, S., and Marcus, S. M. (1998), "Should we adjust for covariates in nonlinear regresson analyses of randomized trials?" *Controlled Clinical Trials*, **19**, 249–256.

Hernán, M. A. (2004), "A definition of causal effect for epidemiological research," *Journal of Epidemiology and Community Health*, **58**, 265–271.

Hernán, M. A. and Robins, J. M. (2006), "Estimating causal effects from epidemiological data," *Journal of Epidemiology and Community Health*, **60**, 578–586.

Hernán, M. A. and Robins, J. M. (2006), "Instruments for Causal Inference. *An Epidemiologist's Dream?*" *Epidemiology*, **17**, 360–372.

Hirano, K. and Imbens, G. W. (2002), "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Heart Catherization," *Health Services and Outcomes Research Methodology*, **2**, 259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, **71**, 1161–1189.

Hoeffding, W. (1952), "The Large Sample Power of Tests Based on Permutations of Observations," *The Annals of Mathematical Statistics*, **23**, 169–192.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, **47**, 663–685.

Hyun, S., Lee, J., and Sun, Y. (2012), "Proportional hazards model for competing risks data with missing cause of failure," *Journal of Statistical Planning and Inference*, **142**, 1767–1779.

Kang, J. D. Y. and Schafer, J. L. (2007a), "Demystifying Double Robustness: a Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, **22**, 523–539.

— (2007b), "Rejoinder: Demystifying Double Robustness: a Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, **22**, 574–580.

Korn, E. L. and Baumrind, S. (1998), "Clinician preferences and the estimation of causal treatment differences," *Statistical Science*, **13**, 209–235.

Kott, P. S. and Liao, D. (2012), "Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-Weighting Routine," *Survey Research Methods*, **6**, 105–111.

Lehmann, E. L. (1951), "Consistency and unbiasedness of certain nonparametric tests," *Annals of Mathematical Statistics*, **22**, 165–179.

— (1963), "Nonparametric Confidence Intervals for a Shift Parameter," *The Annals of Mathematical Statistics*, **34**, 1507–1512.

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistcal Hypotheses*, Springer Texts in Statistics.

Leon, S., Tsiatis, A., and Davidian, M. (2003), "Semiparametric estimation of treatment effect in a pretest-posttest study," *Biometrics*, **59**, 1046–1055.

Lewis, J. A. (1999), "Statistical princples for clinical trials (ICH E9): an introductory note on an international guideline," *Statistics in Medicine*, **18**, 1903–1904.

Loève, M. (1963), *Probability Theory (third edition)*, Springer Verlag, Berlin.

Luenberger, D. G. (1969), *Optimization by Vector Space Methods*, Wiley, New York.

Lumley, T., Shaw, P. A., and Dai, J. Y. (2011), "Connections between survey calibration estimators and semiparametric models for incomplete data," *International Statistical Review*, **79**, 200–220.

Mann, H. and Whitney, D. (1947), "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, **18**, 50–60.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity Score Estimation with Boosted Regression for Evaluating Causal Efects in Observational Studies," *Psychological Methods*, **9**, 403–425.

Moore, K. L. and van der Laan, M. J. (2009), "Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation," *Statistics in Medicine*, **28**, 39–64.

Murphy, S. A., van der Laan, M. J., and Robins, J. M. (2001), "Marginal mean models for dynamic regimes," *Journal of the American Statistical Association*, **96**, 1410–1423.

Newcombe, R. G. (2006), "Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods," *Statistics in Medicine*, **25**, 543–557.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, **5**, 99–135.

Newey, W. K. and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, **4**, 2111–2245.

Nolan, D. and Pollard, D. (1987), "U-Processes: Rates of Convergence," *The Annals of Statistics*, **15**, 780–799.

— (1988), "Functional Limit Theorems for U-Processes," *The Annals of Statistics*, **16**, 1291–1298.

Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012), "Doubly Robust Intrumental Variable Regression," *Statistica Sinica*, **22**, 173–205.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010), "Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content," *The International Journal of Biostatistics*, **6**.

Pierce, D. A. (1982), "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics," *Annals of Statistics*, **10**, 475–478.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002), "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems," *Statistics in Medicine*, **21**, 2917–2930.

Polley, E. and van der Laan, M. (2014), *Super Learner Prediction: the* `SuperLearner` *package*.

Porter, K. E., Gruber, S., van der Laan, M. J., and Sekhon, J. S. (2011), "The Relative Performance of Targeted Maximum Likelihood Estimators," *The International Journal of Biostatistics*, **7**.

Raab, G. M., Day, S., and Sales, J. (2000), "How to select covariates to include in the analysis of a clinical trial," *Controlled Clinical Trials*, **21**, 330–342.

Ridgeway, G. (1999), "The state of boosting," *Computing Science and Statistics*, **31**, 172–181.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. (2015), *Toolkit for Weighting and Analysis of Nonequivalent Groups: the* `twang` *package*.

Ridgeway, G. and McCaffrey, D. F. (2007), "Comment: Demystifying Double Robustness: a Comparison of Alternative Strategies for Estimating a Population Mean for Incomplete Data," *Statistical Science*, **22**, 540–543.

Robins, J. M. (1994), "Correcting for Noncompliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics: Theory and Methods*, **23**, 2379–2412.

— (1998), "Marginal structural model," *1997 Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, 1–10.

— (1999a), "Association, Causation, and Marginal Structural Models," *Synthese*, **121**, 151–179.

— (1999b), "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference," *Statistical Models in Epidemiology: The Environment and Clinical Trials*, **116**, 95–134.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology ," *Epidemiology*, **11**, 550–560.

Robins, J. M., Li, L., Tchetgen Tchetgen, E., and van der Vaart, A. (2008), "Higher Order Influence Functions and Minimax Estimation of Nonliniear Functionals," *IMS Lecture NotesMonograph Series Probability and Statistics Models: Essays in Honor of David A. Freedman*, **2**, 335–421.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992), "Estimating Exposure Effects by Modeling the Expectation of Exposure Conditional on Confounders," *Biometrics*, **48**, 479–495.

Robins, J. M. and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, **16**, 285–319.

Robins, J. M. and Rotnitzky, A. (2001), "Comment on the Bickel and Kwon article, "Inference for semiparametric models: some suestions and an answer"," *Statistica Sinica*, **11**, 920–936.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression-Coefficients when some Regressors are not Always Observed," *Journal of the American Statistical Association*, **89**, 846–866.

Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Comment: Performance of Double-Robust Estimators when Inverse Probability Weights are Highly Variable," *Statistical Science*, **22**, 544–559.

Robinson, L. D. and Jewell, N. P. (1991), "Some surprising results about covariate adjustment in logistic regression models," *International Statistical Review*, **58**, 227–240.

Rosenbaum, P. R. (1984), "Conditional permutation tests and the propensity score in observational studies," *Journal of the American Statistical Association*, **79**, 565–574.

— (2002), "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, **17**, 286–304.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, **70**, 41–55.

Rosenblum, M. and van der Laan, M. J. (2009), "Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models," *Biometrics*, **65**, 937–945.

Rotnitzky, A., Faraggi, D., and Schisterman, E. (2006), "Doubly Robust Estimation of the Area Under the Receiver-Operating Characteristic Curve in the Presence of Verification Bias," *Journal of the American Statistical Association*, **101**, 1276–1288.

Rotnitzky, A., Lei, Q. H., Sued, M., and Robins, J. M. (2012), "Improved Double-Robust Estimation in Missing Data and Causal Inference Models," *Biometrika*, **99**, 439–456.

Rotnitzky, A., Li, L. L., and Li, X. C. (2010), "A Note on Overadjustment in Inverse Probability Weighted Estimation," *Biometrika*, **97**, 997–1001.

Rotnitzky, A. and Robins, J. M. (1997), "Analysis of semi-parametric regression models with non-ignorable non-response," *Statistics in Medicine*, **16**, 81–102.

Rotnitzky, A. and Vansteelandt, S. (2014), "Double-robust methods," in *Handbook of Missing Data Methodology*, eds. Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. A., and Verbeke, G., CRC Press.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, Inc.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, **63**, 581–590.

Rubin, D. B. and van der Laan, M. J. (2008), "Empirical Efficiency Maximization: Improved Locally Efficient Covariate Adjustment in Randomized Experiments and Survival Analysis. ," *International Journal of Biostatistics*, **4**, 1–40.

Särndal, C. E., Swenson, B., and Wretman, J. (1989), "The weighted residual technique for estimating the variance of the general regression estimator," *Biometrika*, **76**, 527–537.

Schafer, J. L. and Graham, J. W. (2002), "Missing Data: Our View of the State of the Art," *Psychological Methods*, **7**, 147–177.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999a), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, **94**, 1096–1120.

— (1999b), "Adjusting For Nonignorable Drop-Out Using Semiparametric Nonresponse Models - Rejoinder," *Journal of the American Statistical Association*, **94**, 1135–1146.

Schisterman, E. and Rotnitzky, A. (2001), "Estimation of the mean of a K-sample U-statistic with missing outcomes and auxiliaries," *Biometrika*, **88**, 713–725.

Senn, S. (2000), "Consensus and controversy in pharmaceutical statistics," *The Statistician*, **49**, 135–176.

— (2006), "Letter to the Editor: "Probabilistic index: an intuitive non-parametric approach to measuring the size of the treatment effects" by L. Acion, J.J. Peterson, S. Temple and S. Arndt," *Statistics in Medicine*, **25**, 3944–3948.

— (2011), "U is for unease: reasons for mistrusting overlap measures for reporting clinical trials," *Statistics in Biopharmaceutical Research*, **3**, 302–309.

— (2012), "Discussion of "Probabilistic Index Models" by O. Thas, J. De Neve, L. Clement and J.P. Ottoy," *Journal of the Royal Statistical Society - Series B*, **74**, 623–571.

Sjölander, A. and Vansteelandt, S. (2011), "Doubly Robust Estimation of Attributable Fractions," *Biostatistics*, **12**, 112–121.

Stefanski, L. A. and Boos, D. D. (2002), "The calculus of M-estimation," *The American Statistician*, **56**, 29–38.

Stephens, A. J., Tchetgen Tchetgen, E. J., and De Gruttola, V. (2013), "Flexible covariate-adjusted exact tests of randomized treatment effects with application to a trial of HIV education," *The Annals of Applied Statistics*, **7**, 2106–2137.

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, **101**, 1619–1637.

— (2007), "Comment: Understanding OR, PS and DR," *Statistical Science*, **22**, 560–568.

— (2010), "Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting," *Biometrika*, **97**, 661–682.

Tan, Z. and Shu, H. (2013), *Improved methods for causal inference and missing data problems: the* `iWeigReg` *package*.

Tchetgen Tchetgen, E. J. and Rotnitzky, A. (2011), "On Protected Estimation of an Odds Ratio Model with Missing Binary Exposure and Confounders," *Biometrika*, **98**, 749–754.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012), "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness, and Sensitivity Analysis," *Annals of Statistics*, **40**, 1816–1845.

Thas, O., De Neve, J., Clement, L., and Ottoy, J. P. (2012), "Probabilistic index models," *Journal of the Royal Statistical Society, Series B-Statistical Methodology*, **74**, 623–671.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, Springer: New York.

Tsiatis, A. A. and Davidian, M. (2007), "Comment: Demystifying Double Robustness: a Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, **22**, 569–573.

Tsiatis, A. A., Davidian, M., and Cao, W. (2011), "Improved Doubly Robust Estimation When Data are Monotonely Coarsened, with Application to Longitudinal Studies with Dropout," *Biometrics*, **67**, 536–545.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008), "Covariate Adjustment for Two-Sample Treatment Comparisons in Randomized Clinical Trials: a Principled yet Flexible Approach," *Statistics in Medicine*, **27**, 4658–4677.

van der Laan, M. J. (2014), "Targeted Estimation of Nuisance Parameters to Obtain Valid Statistical Inference," *International Journal of Biostatistics*, **10**, 29–57.

van der Laan, M. J. and Gruber, S. (2010), "Collaborative Double Robust Targeted Maximum Likelihood Estimation," *International Journal of Biostatistics*, **6**.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, **6**.

van der Laan, M. J. and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, Springer series in Statistics.

van der Laan, M. J. and Rose, S. (2011), *Targeted Learning, Causal Inference for Observational and Experimental Data*, Springer, New York.

van der Laan, M. J. and Rubin, D. B. (2006), "Targeted Maximum Likelihood Learning," *International Journal of Biostatistics*, **2**.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak convergence and empirical processes*, New-York: Springer-Verlag.

Vansteelandt, S. (2012), "Discussion of "Probabilistic Index Models" by O. Thas, J. De Neve, L. Clement and J.P. Ottoy," *Journal of the Royal Statistical Society - Series B*, **74**, 623–571.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012), "On Model Selection and Model Misspecification in Causal Inference," *Statistical Methods in Medical Research*, **21**, 7–30.

Vansteelandt, S. and Keiding, N. (2011), "Invited commentary: G-computation—Lost in translation?" *American Journal of Epidemiology*, **173**, 731–738.

Vansteelandt, S., Vanderweele, T. J., Tchetgen Tchetgen, E. J., and Robins, J. M. (2008), "Multiply Robust Inference for Statistical Interactions," *Journal of the American Statistical Association*, **103**, 1693–1704.

Vermeulen, K., Thas, O., and Vansteelandt, S. (2015), "Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment," *Statistics in Medicine*, **34**, 1012–1030.

Vermeulen, K. and Vansteelandt, S. (2015a), "Bias-Reduced Doubly Robust Estimation," *Journal of the American Statistical Association*, in press.

— (2015b), "Data-Adaptive Bias-Reduced Doubly Robust Estimator," *The International Journal of Biostatistics*, in press.

Wang, L., Rotnitzky, A., and Lin, X. (2010), "Nonparametric regression with missing outcomes using weighted kernel estimating equations," *Journal of the American Statistical Association*, **105**, 1135–1146.

Wilcoxon, F. (1945), "Individual comparisons by ranking methods," *Biometrics*, **1**, 80–83.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Massachusetts.

Wu, P., Han, Y., Chen, T., and Tu, X. M. (2013), "Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics," *Statistics in Medicine*, **33**, 1261–1271.

Yu, Q., Tang, W., Kowalski, J., and Tu, X. M. (2011), "Multivariate U-statistics: a tutorial with applications," *Wiley Interdisciplinary Reviews - Computational Statistics*, **3**, 457–471.

Yu, Z. and van der Laan, M. (2005), "Double Robust Estimation in Longitudinal Marginal Structural Models," *Journal of Statistical Planning and Inference*, **136**, 1061–1089.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008), "Improving efficiency of inferences in randomized clinical trials using auxiliary covariates," *Biometrics*, **64**, 707–715.

Zhang, Z. W., Chen, Z., Troendle, J. F., and Zhang, J. (2012), "Causal inference on quantiles with an obstetric application," *Biometrics*, **68**, 697–706.

Zheng, W. J. and van der Laan, M. J. (2012), "Targeted Maximum Likelihood Estimation of Natural Direct Effects," *International Journal of Biostatistics*, **8**.

Zhou, W. (2008), "Statistical inference for $P(X < Y)$," *Statistics in Medicine*, **27**, 257–279.

Index

— N —

— O —

— P —

— Q —

— R —

— S —

— T —

— U —

— V —

— W —

— Z —