Essays in Econometrics

By

Alexandre Poirier

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James L. Powell, Chair
Professor Bryan S. Graham
Professor Martin Lettau
Professor Demian Pouzo

Spring 2013

Essays in Econometrics

Abstract

Essays in Econometrics

by

Alexandre Poirier

Doctor of Philosophy in Economics

University of California, Berkeley

Professor James L. Powell, Chair

This dissertation consists of two chapters, both contributing to the field of econometrics. The contributions are mostly in the areas of estimation theory, as both chapters develop new estimators and study their properties. They are also both developed for semi-parametric models: models containing both a finite dimensional parameter of interest, as well as infinite dimensional nuisance parameters. In both chapters, we show the estimators' consistency, asymptotic normality and characterize their asymptotic variance. The second chapter is co-authored with professors Jinyong Hahn, Bryan S. Graham and James L. Powell.

In the first chapter, we focus on estimation in a cross-sectional model with independence restrictions, because unconditional or conditional independence restrictions are used in many econometric models to identify their parameters. However, there are few results about efficient estimation procedures for finite-dimensional parameters under these independence restrictions. In this chapter, we compute the efficiency bound for finite-dimensional parameters under independence restrictions, and propose an estimator that is consistent, asymptotically normal and achieves the efficiency bound. The estimator is based on a growing number of zero-covariance conditions that are asymptotically equivalent to the independence restriction. The results are illustrated with four examples: a linear instrumental variables regression model, a semilinear regression model, a semiparametric discrete response model and an instrumental variables regression model with an unknown link function. A Monte-Carlo study is performed to investigate the estimator's small sample properties and give some guidance on when substantial efficiency gains can be made by using the proposed efficient estimator.

In the second chapter, we focus on estimation in a panel data model with correlated random effects and focus on the identification and estimation of various functionals of the random coefficients distributions. In particular, we design estimators for the conditional and unconditional quantiles of the random coefficients distribution. This model allows for irregularly identified panel data models, as in Graham and Powell (2012), where quantiles of the effect are identified by using two subpopulations of "movers" and "stayers", i.e. those for whom the covariates change by a large amount from one period to another, and those for whom covariates remain (nearly) unchanged. We also consider an alternative asymptotic framework where the fraction of stayers in the population is shrinking with the sample size. The purpose of this framework is to approximate a continuous distribution of covariates

where there is an infinitesimal fraction of exact stayers. We also derive the asymptotic variance of the coefficient's distribution in this framework, and we conjecture the form of the asymptotic variance under a continuous distribution of covariates.

The main goal of this dissertation is to expand the choice set of estimators available to applied researchers. In chapter one, the proposed estimator attains the efficiency bound and might allow researchers to gain more precision in estimation, by getting smaller standard errors. In the second chapter, the new estimator allows researchers to estimate quantile effects in a just-identified panel data model, a contribution to the literature.

To Stacy

Contents

Acknowledgments

I thank my advisor Jim Powell for his invaluable guidance, dedication and patience over the years. I am grateful for the amount of time he has taken to advise and support me. I have learned a tremendous amount from my conversations with him.

I am also very thankful to my other dissertation committee members: Bryan Graham, for his constant support, insights and his help in shaping the direction of my research. Demian Pouzo, for his detailed advice and moral support during the dissertation and job market stage. I also thank Martin Lettau for useful comments and advice on research. I am also indebted to Bryan and Jim for having given me the opportunity to do research with them.

I also want to thank Michael Jansson and Denis Nekipelov for useful comments and suggestions. I wish to express a special thanks to Yuriy Gorodnichenko for having been a great and dedicated mentor in my first years as a graduate student. I have also spent much time interacting with fellow graduate students who have helped me in different ways over the years. I especially thank Josh Hausman, Matteo Maggiori, Omar Nayeem, Sebastian Stumpner and James Zuberi. I also acknowledge support from the FQRSC during my graduate studies.

I want to thank my parents and brothers, who have supported me in all my decisions from early in my life all the way to finishing this dissertation. Finally, I thank my amazing fiancee Stacy for being supportive throughout my doctoral studies and for being there during both the good and bad times that come with graduate studies. I dedicate this dissertation to her.

Chapter 1

Efficient Estimation in Models with Independence Restrictions

1.1 Introduction

Many econometric models are identified using zero-covariance or mean-independence restrictions between an unobserved error term and a subset of the explanatory variables. For example, it is common to assume in linear regression models that the error is either uncorrelated with the exogenous variables, or uncorrelated with all functions of the exogenous variables. These restrictions identify the parameters of interest, and efficiency bounds for these type of models have been widely studied, for example in the seminal work of Chamberlain (1987). In other models though, statistical independence of the unobserved error and a subset of the explanatory variables is instead assumed. In most cases, statistical independence is a stronger restriction than mean-independence, as it implies mean-independence of all measurable functions of the unobserved error with respect to the subset of explanatory variables. Independence assumptions are common in recent strands of the literature including in non-linear and non-separable models to allow for heterogenous effects, in the potential outcomes framework and in non-linear structural models.

In this chapter, we compute the efficiency bound for parameters identified by different types of independence restrictions. The efficiency bound for a parameter θ is the smallest possible asymptotic variance for regular asymptotically linear estimators.[1] We use the projection method of Bickel et al. (1993) and Newey (1990c) to compute the bounds. This chapter also derives an estimator that is consistent, asymptotically normal and attains the efficiency bound. We will also highlight the size of the efficiency gains through a Monte Carlo exercise where we evaluate the performance of our estimator.

While imposing mean-independence restrictions (i.e. conditional mean restrictions) is common practice in economics, the stronger independence assumption is useful to consider for two reasons. The first is that some models require statistical independence for identification purposes, as is sometimes the case in non-linear semiparametric and nonparametric models. The second reason is that even if mean-independence restrictions are sufficient to identify

[1] See Bickel et al. (1993) for the formal definition.

the parameter of interest, imposing independence can be justified by economic conditions. The additional information included in the independence restriction can potentially be used to derive estimators with smaller asymptotic variances. In either case, using an efficient estimator will ensure that the estimator's large-sample properties cannot be improved on.

We perform efficiency bound calculations for general classes of model where a residual function ($Y - W\theta$ for example) is independent (or conditionally independent) of an exogenous variables (X). We also allow for the presence of an unknown function in the residual function, and compute the semiparametric efficiency bound for the finite dimensional parameter in that case. The estimator proposed uses a framework similar to that of efficient GMM with two differences. The first is that we use covariance restrictions rather than moment restrictions, which leads to a different optimal weighting matrix. The estimation procedure is based on an increasing number of zero-covariance conditions between some functions of the error and the exogenous variables. Second, the number of restrictions is growing with the sample size and we derive maximal rates at which that number grows to infinity. We will show that by letting the functions considered be in specific classes, independence will be asymptotically equivalent to the zero-covariance conditions, as their number increases. We will further characterize results when the class of function chosen are complex exponential functions, by using facts about characteristic functions.

1.1.1 Related Literature

Efficiency bound calculations for mean-independence restrictions were performed in Chamberlain (1987) and Bickel et al. (1993), and efficient estimators were developed in Newey (1990b) and Newey (1993). For models with the stronger unconditional independence restrictions, early results can be found in MaCurdy (1982), Newey (1989) and Newey (1990a). MaCurdy (1982) shows that using zero-covariance restrictions between higher moments yields asymptotic efficiency improvements. Both Newey (1989) and Newey (1990a) propose an estimator that minimizes a V-statistic based on an approximation to the efficient score. Newey (1989) constructs a locally efficient estimator, meaning that efficiency is achieved if one correctly postulates a parametric family for the unobserved error's distribution, while the estimator in Newey (1990a) is globally efficient but requires additional assumptions since it nonparametrically estimates the distribution of the error in a first-stage estimate. Hansen et al. (2010) consider a linear instrumental variables system with the instruments independent from the error term and propose a locally efficient estimator. By contrast, the estimator we propose is globally efficient, and is obtained by minimizing a GMM objective function with an optimal weighting matrix. Manski (1983) also proposed the "closest empirical distribution" approach, and Brown and Wegkamp (2002) derived asymptotic properties of estimators based on this approach. Though not considered in this chapter, empirical likelihood estimators, such as those proposed in Donald et al. (2003) and Donald et al. (2008) for mean-independence restrictions, can also attain efficiency bounds and often exhibit better small-sample properties than the corresponding two-step GMM estimator.

An alternative approach for deriving an efficient estimator was proposed in Carrasco and Florens (2000). Their estimator is based on the estimation of a method of moments estimator

with an uncountable number of moment restrictions. This CGMM (Continuum of GMM) estimator will be efficient when the optimal weighting "operator" is used to optimally reweigh the continuum of moment conditions. This objective function is computationally challenging to evaluate. More generally, this problem suffers from the ill-posed inverse problem, and a Tikhonov regularization is suggested. By comparison, the estimator we propose uses a finite but growing number of zero-covariance conditions such that asymptotically the continuum of covariance restrictions are used.

For models defined by conditional independence restrictions, the literature has mostly focused on testing rather than estimation. Su and White (2008) propose a nonparametric test of conditional independence which uses the Hellinger distance between two conditional densities, and Linton and Gozalo (1996) propose a test based on joint probabilities on half-spaces. It is interesting to note that single-index restrictions are equivalent to a conditional independence restriction, as in Klein and Spady (1993) or Ichimura (1993). Cosslett (1987) has computed the efficiency bound for the semiparametric binary choice model with $X \perp \epsilon$, and Klein and Spady (1993) have derived an efficient estimator under a single index restriction in the same binary choice model by proposing a semiparametric maximum likelihood estimator. Lee (1995) derived an efficient estimator for the multiple response model under distributional single-index restrictions, also a conditional independence restriction. These additional restrictions are useful for estimating transformation models as in Han (1987), which include ordered choice models and semiparametric censored regression models (e.g. Powell (1984)).

Finally, Ai and Chen (2003) made a significant contribution by deriving a consistent and efficient estimator for models that satisfy a mean-independence restriction with an unknown finite dimensional parameter (θ_0) and an unknown infinite dimensional nuisance function ($F_0(\cdot)$):

$$E[\rho(Y, W, \theta_0, F_0(\cdot))|X] = 0.$$

The efficiency bound for θ_0 under $\rho(Y, W, \theta_0, F_0(\cdot)) \perp X$ has not been studied so far. Komunjer and Santos (2010) examine a simple semiparametric BLP model where the finite-dimensional parameter of interest is identified through an independence restriction. Using a Cramer-Von-Mises objective function, they derive a consistent and asymptotically normal estimator. Santos (2011) suggests an estimator in more general semiparametric non-separable transformation models with independence restrictions. He does not establish efficiency of his estimator and instead shows that the class of transformation models he investigates is not regular since his model is not differentiable in quadratic mean at the true parameter value.

1.1.2 Model

The different econometric models with independence restrictions will be categorized below. We first consider a general model with a conditional independence restriction:

$$\rho(Y, W, \theta_0) = \epsilon \text{ with } \epsilon \perp X|Z \tag{1.1}$$

where $\rho(\cdot)$ is a known residual function, and θ is a finite dimensional parameter with $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $d_\theta = \dim(\theta)$. Also, $W \in \mathcal{W} \subset \mathbb{R}^{d_W}$, $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$, $Z \in \mathcal{Z} \subset \mathbb{R}^{d_Z}$, $Y \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ and $\epsilon \in \mathcal{E} \subset \mathbb{R}^{d_\epsilon}$. The joint distribution of (Y, W, ϵ, X, Z) is unknown and is a member of the class of distributions which satisfy (1.1). This is a semiparametric model since that class of distributions is infinite dimensional and the parameter of interest θ_0 is finite dimensional. The conditional independence restriction can be expressed as a restriction on the conditional CDF of X and ϵ given Z:

$$P(\rho(Y, W, \theta_0) \leq e, X \leq x|Z = z) = P(\rho(Y, W, \theta_0) \leq e|Z = z)P(X \leq x|Z = z)$$

for all $e \in \mathcal{E}$, $x \in \mathcal{X}$ and $z \in \mathcal{Z}$.[2] Note that model (1.1) generalizes unconditional independence restrictions, since when Z is a constant, the conditional independence restriction is also an unconditional independence restriction:[3]

$$\rho(Y, W, \theta_0) = \epsilon \text{ with } \epsilon \perp X. \tag{1.2}$$

Here are some examples included in model (1.1):

Example 1.1.1 (Linear Regression) *Let $W = X$, and $Y = X\theta_0 + \epsilon$ with $\epsilon \perp X$. This is a linear regression model with independent errors, studied in Bickel et al. (1993). The residual function is $\rho(Y, X, \theta) = Y - X\theta$.*

Example 1.1.2 (Linear Instrumental Variables Regression) *Let $Y = W\theta_0 + \epsilon$ and $\epsilon \perp X$. This is a linear IV model with independent errors. This is a stronger restriction than $E[\epsilon|X] = E[\epsilon]$ or $Cov(\epsilon, X) = 0$, which are typically assumed in IV models. The residual function is $\rho(Y, W, \theta) = Y - W\theta$.*

Example 1.1.3 (Semilinear Regression) *Let $Y = W\theta_0 + G_0(Z, U)$ with $U \perp (W, Z)$, $G_0(\cdot, \cdot)$ an unknown function, and U unobservable. This is a generalization of Robinson (1988) semilinear regression model which allows the marginal effect of Z on Y to vary across observationally equivalent units. Let $\rho(Y, W, \theta) = Y - W\theta$ and let $\epsilon = G_0(Z, U)$. We will show later on that this model can be represented by $\rho(Y, W, \theta) \perp W|Z$, thus fitting model (1.1).*

Example 1.1.4 (Potential Outcomes) *In the potential outcomes literature, it is commonly assumed that $(Y_0, Y_1) \perp D|X$ where (Y_0, Y_1) are the outcomes when the unit is untreated and treated, respectively, D is the treatment dummy and X are covariates. Let $Y = DY_1 + (1 - D)Y_0$ be the observed outcome. Let $Y = m(D, X, \epsilon, \theta_0)$, and let $m(\cdot)$ be invertible in ϵ, and let $\rho(Y, D, X, \theta_0)$ be its inverse. We can show that this model is equivalent to $\rho(Y, D, X, \theta) \perp D|X$.*

[2] For alternative characterizations of conditional independence, see Dawid (1979).

[3] Conditioning on a constant being equal to itself is exactly equivalent to not conditioning.

Alternatively, other economic models can be represented with the following conditional independence restriction:

$$Y \perp X | V(X, \theta_0). \tag{1.3}$$

with $V(\cdot)$ an index function mapping to $\mathbb{R}^{d_V}$. We will often use $V(X, \theta_0) = X'\theta_0$ and therefore $d_V = 1$. This model is similar to (1.1), but instead the parameter of interest θ_0 appears in the conditioning variable. This model is used in many examples in discrete choice analysis and in non-invertible models, for example:

Example 1.1.5 (Semiparametric Binary Response) *Let $Y = \mathbf{1}(\epsilon \leq X_1 + X_2\theta_0)$, and let this single-index restriction hold: $E[Y|X] = E[Y|X'\theta_0]$ where $X'\theta_0 = X_1 + X_2\theta_0$.*[4] *We can show that this single index restriction is equivalent to $Y \perp X | X'\theta_0$.*

Example 1.1.6 (Censored Regression) *Let $Y = \max\{X'\theta_0 + \epsilon, 0\}$, and let $\epsilon \perp X | X'\theta_0$. Then, this restriction implies that $Y \perp X | X'\theta_0$.*

We also consider a generalization of model (1.2) where the residual function includes an unknown function $F_0(\cdot)$ (i.e. a nuisance function):

$$\rho(Y, W, \theta_0, F_0(\cdot)) = \epsilon \text{ with } \epsilon \perp X. \tag{1.4}$$

Here is an example of a model which satisfies (1.4):

Example 1.1.7 (Transformation IV) *Let $Y = \Lambda_0(W\theta_0 + \epsilon)$ and $\epsilon \perp X$, an instrument. This is a transformation IV model, and we assume that $\Lambda_0(\cdot)$ is an unknown and strictly increasing function. Let $F_0(\cdot) = \Lambda_0^{-1}(\cdot)$. We can then let $\rho(Y, W, \theta_0, F_0(\cdot)) = F_0(Y) - W\theta_0$, and therefore this is an example which satisfies restriction* (1.4).

Example 1.1.8 (Semilinear Regression 2) *Let $Y = W\theta_0 + g_0(Z) + \epsilon$ and $\epsilon \perp (W, Z)$. This is Robinson (1988)'s model with independent errors. We can then let $\rho(Y, W, \theta_0, F_0(\cdot)) = Y - W\theta_0 - F_0(Z)$, $X = (W, Z)$ and therefore this is also an example which satisfies restriction* (1.4).

1.1.3 Estimation Technique

For model (1.2), the efficient estimator we propose relies on zero-covariance restrictions between the functions $e^{is\rho(Y,W,\theta_0)}$ and e^{itX} for any values of s and t. We show that the GMM-type estimator of an increasing number of these covariance restrictions with an optimal weighting matrix attains the efficiency bound. This estimator makes full use of the independence restriction through the complex exponential approximating functions, and efficiency is attained by using a GMM-type setup with an optimal weighting matrix. To get a concrete idea of how the estimator is derived, we will present the population objective function. Let

[4] For more on single-index restrictions see Ichimura (1993), Powell et al. (1989) and Bester and Hansen (2009).

$g(s,t,\theta) = \text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}\right)$, and let Ω be a linear operator on functions from $\mathbb{R}^2$ to $\mathbb{R}^2$ with kernel $k(s,t,s',t') = \text{Cov}\left(e^{is\epsilon}, e^{is'\epsilon}\right)\text{Cov}\left(e^{itX}, e^{it'X}\right)$. The population criterion function is equal to:

$$R(\theta) = \|g(\cdot,\cdot,\theta)\|_\Omega^2 \tag{1.5}$$

where $\|g\|_\Omega^2 = (\Omega^{-1/2}g, \Omega^{-1/2}g)$ denotes the norm of g in the reproducing kernel Hilbert space (RHKS) defined by Ω.[5] We approximate the population objective function by a sieve. Let $\hat{\bar{g}}^{KL}(\theta)$ be a KL by 1 vector of sample covariances between $e^{is\rho(Y,W,\theta)}$ and e^{itX} evaluated at $K \times L$ different values of (s,t). We also let $\widehat{\bar{\Omega}}^{KL}$ be a $KL \times KL$ dimensional matrix which is a finite-dimensional approximation of the operator Ω. We will show that $\widehat{\bar{\Omega}}^{KL}$ is an optimal weighting matrix for the sample covariances $\hat{\bar{g}}(\theta)$, and that the estimator:

$$\widehat{\theta} = \text{argmin}_{\theta\in\Theta}\widehat{R}^{KL}(\theta)$$
$$\widehat{R}^{KL}(\theta) = \hat{\bar{g}}(\theta)^{KL\prime}\left(\widehat{\bar{\Omega}}^{KL}\right)^{-1}\hat{\bar{g}}(\theta)^{KL} = \|\hat{\bar{g}}\|^2_{\widehat{\bar{\Omega}}^{KL}}$$

will attain the efficiency bound as K and $L \to \infty$, and other regularity conditions are satisfied.

This chapter is organized as follows. Section 2 derives the efficiency bound for the different restrictions considered. Sections 3 introduces efficient estimators for parameters in models represented by (1.1). Section 4 conjectures an efficient estimator for θ_0 in models with independence restrictions and unknown functions. A Monte Carlo study is presented in section 5, and section 6 concludes.

1.2 Computation of Efficiency Bounds

A practical method for computing efficiency bounds in regular semiparametric models is the projection method, pioneered in Bickel et al. (1993) and surveyed in Newey (1990c) and Tsiatis (2006). Using this method entails devising a generic parametric submodel of the semiparametric model, and computing the Cramer-Rao lower bound for the estimation of θ_0 given this submodel. The efficiency bound of θ_0 in the parametric submodel cannot be larger than the efficiency bound in the larger semiparametric model, since the submodel imposes restrictions in addition to those in the semiparametric model. The efficiency bound will be equal to the supremum of the Cramer-Rao bounds for all submodels, if the supremum exists. We project the score function of the semiparametric model on the nuisance tangent space, yielding the efficient score, which is sufficient for computing the efficient influence function and the semiparametric efficiency bound of the model.

[5] See Parzen (1959) for an introduction to reproducing kernel Hilbert spaces.

1.2.1 Unconditional Independence

To compute the efficiency bound of θ_0 in model (1.2), the following assumptions will suffice:

Assumption 1.2.1 (UI)

(a) *(Parameter Space)* $d_\theta = 1$ *and* Θ *is compact;*

(b) *(Identification)* $\rho(Y, W, \theta) \perp X \Rightarrow \theta = \theta_0$*;*

(c) *(Invertibility)* $d_Y = d_\epsilon = 1$ *and* $\rho(\cdot, W, \theta)$ *is invertible for all* $\theta \in \Theta$ *a.s. -* W*;*

(d) *(Finite Fisher Information)* ϵ *has a continuously differentiable density function* $f_\epsilon(\cdot)$ *and* $0 < E\left[\dfrac{f'_\epsilon(\epsilon)^2}{f_\epsilon(\epsilon)^2}\right] < \infty$*;*

(e) *(Differentiability)* $\rho(Y, W, \theta)$ *is differentiable with respect to* θ *a.s. -* (Y, W)*;*

(f) *(Finite Second Moments of Derivative)* $E[\|\rho_\theta(Y, W, \theta_0)\|^2] < \infty$.

In assumption UI(a) we let the parameter space be one-dimensional. This is done to simplify calculations, and is without great loss of generality. Assumption UI(b) is a high level assumption, and identification of θ_0 will be taken for granted in this chapter. Typically, since θ_0 is a finite dimensional parameter, it will be overidentified because independence restrictions are equivalent to an infinite number of covariance restrictions. Assumption UI(c) further restricts our attention to models where there exists a bijection between scalars ϵ and Y. Most of the results can be generalized to $d_Y = d_\epsilon > 1$ straightforwardly, and this assumption is done to ease calculations. Many models satisfy this restriction, such as most linear, non-linear and quantile regression models. This restriction does rule out many discrete Y models, such as binary response $Y = \mathbf{1}(\epsilon \leq X'\theta_0)$, but some of these models will satisfy an independence restriction as in (1.3). Assumption UI(d) is substantially different from the usual $E[\epsilon^2] < \infty$ assumed for models with mean-independence restrictions or covariance restrictions, since for many distributions with infinite variances (e.g. Cauchy, Student with 2 degrees of freedom), the location score $\dfrac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ is bounded, thus its Fisher information exists. Therefore, independence restrictions can accommodate more models with thick-tailed error distributions than those with moment-based restrictions. Finally, assumptions UI(e)-(f) are regularity conditions necessary to insure the existence of the semiparametric efficiency bound and will be discussed further in the context of an example.

Lemma 1.2.2 *Under assumption* (1.2.1) *the efficient score of the model identified by* $\rho(Y, W, \theta) \perp X$ *is:*

$$
\begin{aligned}
S^{\mathit{eff}}(X,\epsilon,\theta_0) = & E\left[J(Y,W,\theta_0)|X,\epsilon\right] - E\left[J(Y,W,\theta_0)|\epsilon\right] \\
& + \frac{\partial}{\partial \epsilon}(E[\rho_\theta(Y,W,\theta_0)|X,\epsilon] - E[\rho_\theta(Y,W,\theta_0)|\epsilon]) \\
& + \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}(E[\rho_\theta(Y,W,\theta_0)|X,\epsilon] - E[\rho_\theta(Y,W,\theta_0)|\epsilon])
\end{aligned}
$$

where $J(Y,W,\theta_0) = \dfrac{\partial}{\partial \theta}\log\left|\dfrac{\partial}{\partial y'}\rho(Y,W,\theta)\right|_{\theta=\theta_0}$ *is the Jacobian. The efficient influence function is:*

$$
\psi^{\mathit{eff}}(X,\epsilon,\theta_0) = \left(V^{\mathit{eff}}(\theta_0)\right)^{-1} S^{\mathit{eff}}(X,\epsilon,\theta_0)
$$

where $V^{\mathit{eff}}(\theta_0)$ *is the semiparametric efficiency bound:*

$$
\begin{aligned}
V^{\mathit{eff}}(\theta_0)^{-1} = & \; Var\left[E\left[J(Y,W,\theta_0)|X,\epsilon\right] - E\left[J(Y,W,\theta_0)|\epsilon\right]\right] \\
& + 2Cov\left(E\left[J(Y,W,\theta_0)|X,\epsilon\right] - E\left[J(Y,W,\theta_0)|\epsilon\right], \frac{\partial}{\partial \epsilon}h(X,\epsilon) + \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}h(X,\epsilon)\right) \\
& + Var\left[\frac{\partial}{\partial \epsilon}h(X,\epsilon) + \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}h(X,\epsilon)\right]
\end{aligned}
$$

with $h(X,\epsilon) = E[\rho_\theta(Y,W,\theta_0)|X,\epsilon] - E[\rho_\theta(Y,W,\theta_0)|\epsilon]$.

Discussion of Lemma and Linear Regression Example

To gain an insight on the efficient score and variance in this model, it is useful to compare them to their counterparts in the mean-independence model. Independence of ϵ and X will still hold, but this information will not be used in the computation of the efficient score function. For simplicity, we will assume that the Jacobian term is constant, and therefore will not appear in the efficiency calculations. In the mean-independence case $E[\rho(Y,W,\theta_0)|X] = E[\rho(Y,W,\theta_0)]$, the efficient score and variance bound are:

$$S_1^{\text{eff}}(X, \epsilon, \theta_0) = (E[\rho_\theta(Y, W, \theta_0)|X] - E[\rho_\theta(Y, W, \theta_0)]) \frac{\epsilon - E[\epsilon]}{\text{Var}[\epsilon]}$$
$$V_1^{\text{eff}}(\theta_0) = \frac{\text{Var}[\epsilon]}{\text{Var}[E[\rho_\theta(Y, W, \theta_0)|X]]}.$$

We can use the Cauchy-Schwarz inequality to show that $V_1^{\text{eff}}(\theta_0) \geq V^{\text{eff}}(\theta_0)$. These results fit in the optimal instrument framework, and the optimal instrument here is equal to $E[\rho_\theta(Y, W, \theta_0)|X] - E[\rho_\theta(Y, W, \theta_0)]$. The main difference is the presence of the location score $\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ and the conditioning of $\rho_\theta(Y, W, \theta_0)$ on both X and ϵ, rather than only on X. We can see that when $E[\rho_\theta(Y, W, \theta_0)|X, \epsilon] - E[\rho_\theta(Y, W, \theta_0)|\epsilon]$ does not depend on ϵ, and when the location score is equal to $\frac{\epsilon - E[\epsilon]}{\text{Var}[\epsilon]}$ the two efficient scores will be equal.

To illustrate the efficiency gains, consider the endogenous linear regression model in example (1.1.2): $\rho(Y, W, \theta_0) = Y - W\theta_0 = \epsilon$ with ϵ independent from X, the instrument. Assumption UI(e) is trivially satisfied, and assumption UI(f) is satisfied if we assume that W has a finite variance. The existence of moments of X is not required for the computation of the efficiency bound. The efficient score under independence is $S^{\text{eff}}(X, \epsilon, \theta_0) = \frac{\partial}{\partial \epsilon} h(X, \epsilon) + \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)} h(X, \epsilon)$ where $h(X, \epsilon) = -E[W|X, \epsilon] + E[W|\epsilon]$. When (W, X, ϵ) are trivariate normal, we have that $E[W|X, \epsilon]$ is a linear function of X and ϵ, so that $E[W|X, \epsilon] = aX + b\epsilon$ for some a and b. Therefore, $h(X, \epsilon) = -a(X - E[X])$, and the term $\frac{\partial}{\partial \epsilon} h(X, \epsilon)$ will be equal to zero. Also, when the distribution of ϵ is normal with mean μ and variance σ^2, the location score $\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ is equal to $\frac{\mu - \epsilon}{\sigma^2}$, and the inverse of its variance is σ^2. This implies that independence is equivalent to mean-independence when (Z, X, ϵ) are jointly normally distributed. This equivalence is a natural consequence of the fact that jointly normal distributions are uncorrelated if and only if they are independent.

From looking at the differences between these efficient scores, we expect efficiency gains when: **(1)** the distribution of ϵ is not normal, and **(2)** when the conditional expectation $E[\rho_\theta(Y, W, \theta_0)|X, \epsilon]$ is nonlinear in X. In the exogenous linear regression model of example (1.1.1) ($\rho(Y, X, \theta_0) = Y - X\theta_0 \perp X$) the only efficiency gains will come from the non-normality of ϵ, and the ratio of the efficiency bounds is:

$$\frac{V_1^{\text{eff}}(\theta_0)}{V^{\text{eff}}(\theta_0)} = \text{Var}[\epsilon] \, \text{Var}\left[\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}\right].$$

When this ratio is large, using the additional information in the independence restriction will yield large efficiency gains compared to using only the mean-independence restriction.

To illustrate an alternative source of efficiency gains from using independence, consider a pathological linear IV regression example. Let

$$Y = W\theta_0 + \epsilon,$$

and

$$W = X^2\epsilon + \eta.$$

X is the instrument, and (X, ϵ, η) are jointly distributed along a trivariate normal distribution with mean 0 and identity variance matrix. The two-stage least squares estimator will likely have poor properties, since $E[WX] = 0$. In fact, we have that $E[W|X] = 0$, meaning that X provides no useful information on the first moment of W. But it is clear that knowledge of X should help in predicting W, since they are not independent. The efficiency bound for this model under the mean-independence restriction $E[\epsilon|X] = E[\epsilon]$ is infinite, therefore we cannot use it to derive a $\sqrt{N}$-consistent estimator. In this model, it is useful to consider stronger versions of the usual 2SLS assumption, specifically we instead work with the condition $\epsilon \perp X$. In this case, the efficiency bound of θ_0 is equal to $\operatorname{Var}[\epsilon^2]^{-1} \operatorname{Var}[X^2]^{-1}$, so identification of θ_0 is regular despite the fact that linear IV regression techniques cannot recover θ_0.[6] This result can be connected to the literature on weak instruments, and shows that that using covariance between non-linear functions of ϵ and Z as a basis for estimation relaxes the need for $\operatorname{Cov}(W, X) \neq 0$, as long as W and X are not independent. This example is somewhat extreme, but we can see that if $\operatorname{Cov}(W, X)$ is small, but W and X depend non-linearly, efficiency gains from using independence can be very large.

An important point is that throughout this discussion we have assumed that $\epsilon \perp X$. Using an estimator that assumes full independence when mean-independence holds, but independence does not hold will likely lead to inconsistent estimates for θ_0 and invalid standard errors. For example, let

$$Y = X\theta_0 + \epsilon$$

and

$$\epsilon = XU,$$

where $U \perp X$ and $E[U] = 0$. This is a linear regression model with $E[\epsilon X] = E[X^2]E[U] = 0$, and multiplicative heteroskedasticity. Since $E[\epsilon^2|X] = E[U^2]X^2$, the variables ϵ and X are not independent. θ_0 is identified, since we have $E[\epsilon X] = 0$, but we do not have $\epsilon \perp X$, and therefore,

$$Y - X\theta \perp X$$

yields

$$X(U + \theta_0 - \theta) \perp X,$$

and imposing $\theta = \theta_0$ does not satisfy this relationship. In fact, unless U is a constant, no

[6] Since we assumed normality, higher moments of X and ϵ all exist.

value of θ will satisfy this relationship, and therefore estimators based on this identifying restriction will not be asymptotically consistent.

1.2.2 Conditional Independence

To compute the efficiency bound of θ_0 in model (1.1), the following assumptions will suffice:

Assumption 1.2.3 (CI)

(a) *(Parameter Space)* $d_\theta = 1$ *and* Θ *is compact;*

(b) *(Identification)* $\rho(Y, W, \theta) \perp X | Z \Rightarrow \theta = \theta_0$*;*

(c) *(Invertibility)* $d_Y = d_\epsilon = 1$ *and* $\rho(\cdot, W, \theta)$ *is invertible for all* $\theta \in \Theta$ *a.s. -* W*;*

(d) *(Finite Conditional Fisher Information)* ϵ*'s conditional density function* $f_{\epsilon|Z}(\cdot|\cdot)$ *is continuously differentiable in* ϵ *and* $0 < E\left[\dfrac{f'_{\epsilon|Z}(\epsilon|Z)^2}{f_{\epsilon|Z}(\epsilon|Z)^2}\right] < \infty$*;*[7]

(e) *(Differentiability)* $\rho(Y, W, \theta)$ *is differentiable with respect to* θ *a.s. -* (Y, W)*;*

(f) *(Finite Second Moments of Derivative)* $E[\|\rho_\theta(Y, W, \theta_0)\|^2] < \infty$*.*

The assumptions required for lemma (1.2.4) are very similar to assumptions UI(a)-(f). The main difference is in assumption CI(d), where the location score $\dfrac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ is replaced by the conditional location score $\dfrac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}$. This model nests the unconditional independence model since we can let Z be a constant random variable, which will make all assumptions in (1.2.3) equivalent to those in (1.2.1).

Lemma 1.2.4 *Under assumptions* (1.2.3) *the efficient score of the model identified by* $\rho(Y, W, \theta) \perp X | Z$ *is:*

$$\begin{aligned} S^{eff}(X, \epsilon, Z, \theta_0) = {} & E\left[J(Y, W, \theta_0)|X, \epsilon, Z\right] - E\left[J(Y, W, \theta_0)|\epsilon, Z\right] \\ & + \frac{\partial}{\partial \epsilon}(E[\rho_\theta(Y, W, \theta_0)|X, \epsilon, Z] - E[\rho_\theta(Y, W, \theta_0)|\epsilon, Z]) \\ & + \frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}(E[\rho_\theta(Y, W, \theta_0)|X, \epsilon, Z] - E[\rho_\theta(Y, W, \theta_0)|\epsilon, Z]) \end{aligned}$$

[7] $f'_{\epsilon|Z}(\epsilon|Z)$ denotes the partial derivative of $f_{\epsilon|Z}(\epsilon|Z)$ with respect to ϵ.

where $J(Y,W,\theta_0) = \dfrac{\partial}{\partial\theta}\log\left|\dfrac{\partial}{\partial y'}\rho(Y,W,\theta)\right|_{\theta=\theta_0}$ *is the Jacobian. The efficient influence function is:*

$$\psi^{eff}(X,\epsilon,Z,\theta_0) = \left(V^{eff}(\theta_0)\right)^{-1} S^{eff}(X,\epsilon,Z,\theta_0)$$

where $V^{eff}(\theta_0)$ *is the semiparametric efficiency bound:*

$$\begin{aligned} V^{eff}(\theta_0)^{-1} = {} & Var\left[E\left[J(Y,W,\theta_0)|X,\epsilon,Z\right] - E\left[J(Y,W,\theta_0)|\epsilon,Z\right]\right] \\ & +2cov\left(E\left[J(Y,W,\theta_0)|X,\epsilon,Z\right] - E\left[J(Y,W,\theta_0)|\epsilon,Z\right], \frac{\partial}{\partial\epsilon}h(X,\epsilon,Z)\right. \\ & \left. +\frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}h(X,\epsilon,Z)\right) + Var\left[\frac{\partial}{\partial\epsilon}h(X,\epsilon,Z) + \frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}h(X,\epsilon,Z)\right] \end{aligned}$$

with $h(X,\epsilon,Z) = E[\rho_\theta(Y,W,\theta_0)|X,\epsilon,Z] - E[\rho_\theta(Y,W,\theta_0)|\epsilon,Z]$.

Discussion of Lemma and Semilinear Regression Example

Conditional independence is a stronger restriction than the following conditional mean-independence restriction:

$$E[\rho(Y,W,\theta_0)|X,Z] = E[\rho(Y,W,\theta_0)|Z].$$

Assuming we have a model where the Jacobian term in the efficient score disappears (i.e. $E\left[J(Y,W,\theta_0)|X,\epsilon,Z\right] - E\left[J(Y,W,\theta_0)|\epsilon,Z\right] = 0$), we can compute the efficient score under the mean-independence restriction as in Chamberlain (1987), which is:

$$\begin{aligned} S_1^{\text{eff}}(X,\epsilon,Z,\theta_0) &= (E[\rho_\theta(Y,W,\theta_0)|X,Z] - E[\rho_\theta(Y,W,\theta_0)|Z])\frac{\epsilon - E[\epsilon|Z]}{E[\text{Var}\left[\epsilon|Z\right]]} \\ V_1^{\text{eff}}(\theta_0) &= \frac{E[\text{Var}\left[\epsilon|Z\right]]^2}{E[\text{Var}\left[E[\rho_\theta(Y,W,\theta_0)|X,Z]|Z\right]\text{Var}\left[\epsilon|Z\right]]}. \end{aligned}$$

If conditional mean-independence restrictions also identify θ_0, we can compare the efficiency bounds $V^{\text{eff}}(\theta_0)$ and $V_1^{\text{eff}}(\theta_0)$. Comparing their efficient scores, we can see that the

efficiency gains are larger when the data generating process has the following features: (1) the conditional location score $\frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}$ differs greatly from $\frac{\epsilon - E[\epsilon|Z]}{E[\text{Var}[\epsilon|Z]]}$, that is, it is non-normal; or (2), $E[\rho_\theta(Y, W, \theta_0)|X, \epsilon, Z]$ depends non-linearly on X or Z. These conditions are substantially similar to those in the unconditional independence case, and may offer some guidance as to whether imposing conditional independence of the error will yield large efficiency improvements.

In some cases, weakening the conditional independence assumption will result in a loss of identification. Consider the semi-linear regression model $Y = W\theta_0 + G_0(Z, U)$ with $U \perp (W, Z)$, a generalization of Robinson (1988)'s regression model also proposed in Santos (2011). This model allows for heterogeneous effects of Z on Y through $\frac{\partial}{\partial Z} G_0(Z, U)$ since $G_0(\cdot, \cdot)$ is unknown and potentially non-linear. This is a form of unobserved heterogeneity, a topic of great importance for microeconomic data, for example in the treatment effects literature.[8] The restriction $E[U|W, Z] = E[U|Z]$ cannot be used to identify θ_0 unless we we assume that $G_0(Z, U) = g_0(Z) + U$, as in Robinson (1988). Let $\epsilon = G_0(Z, U)$, and consider the following set of additional assumptions:

Assumption 1.2.5 (SL)

(a) *(Strict Monotonicity)* $G_0(z, \cdot)$ *is invertible for every value* $z \in \mathcal{Z}$ *a.s.;*

(b) *(Conditional Independence)* $U \perp W|Z$*;*

(c) *(Unconditional Independence)* $U \perp (W, Z)$*.*

Lemma 1.2.6 *Under assumptions* (1.2.3)*(c)-(f) and either assumptions* (1.2.5) *(a) and (b) or assumptions* (1.2.5) *(a) and (c),* θ_0 *is identified and its efficiency bound is:*

$$V^{eff}(\theta_0) = E\left[\left(\frac{f'_{\epsilon|Z}(G_0(Z, U)|Z)}{f_{\epsilon|Z}(G_0(Z, U)|Z)}\right)^2 (W - E[W|Z])^2\right]^{-1}.$$

This lemma shows that this model's finite-dimensional parameter θ_0 is identified, and also that the efficiency bound does not depend on whether we make the unconditional or conditional independence assumption. Intuitively, the unconditional independence restriction SL(c) implies the conditional independence restriction SL(b) and also that $U \perp Z$, but both U and Z only appear inside an unrestricted and potentially non-linear function $G_0(\cdot, \cdot)$. Because of this, U can be normalized in a way that lets it be independent from Z without altering the other model assumptions. This bound will differ from the one in Robinson (1988)

[8] See Chesher (2003), Evdokimov (2009) and Imbens and Newey (2009) for example.

because he only uses mean-independence rather than independence. If it is indeed true that $G_0(Z,U) = g_0(Z) + U$ the efficiency bound becomes

$$V^{\text{eff}}(\theta_0) = E\left[\left(\frac{f_U'(U)}{f_U(U)}\right)^2\right]^{-1} E\left[\text{Var}\left[W|Z\right]\right]^{-1},$$

which is the bound computed in Bhattacharya and Zhao (1997) for Robinson's model with $U \perp (W,Z)$. They compute the bound and derive an estimator which attains it, thereby making the existence of Var $[U]$ unnecessary for estimation, since a finite Fisher information for U will suffice. Our model places fewer restrictions on the unknown function, but the efficiency bounds are the same in this special case, therefore it is more general. Since $G_0(z,\cdot)$ was assumed invertible, we can normalize the distribution of U to be Uniform$[0,1]$ and get that $G_0(Z,\tau)$ is quantile τ of the non-parametric component.

1.2.3 Conditional Independence with Parameter in the Conditioning Variable

Denote $V_\theta(X,\theta_0) = \dfrac{\partial}{\partial\theta}V(X,\theta)|_{\theta=\theta_0}$. To compute the efficiency bound of θ_0 in model (1.3), the following assumptions will suffice:

Assumption 1.2.7 (CI2)

(a) *(Parameter Space)* $d_\theta = 1$ *and* Θ *is compact;*

(b) *(Identification)* $Y \perp X | V(X,\theta) \Rightarrow \theta = \theta_0$*;*

(c) *(Smoothness) The conditional density of* Y *given* $V = v$ *is continuously differentiable in* v*;*

(d) *(Differentiability)* $V(X,\theta)$ *is differentiable with respect to* θ *a.s. -* X*;*

(e) *(Finite Second Moments of Derivative)* $E\left[\|V_\theta(X,\theta_0)\|^2\right] < \infty$.

Assumption CI2(a) and CI2(b) are similar to those assumed in previous efficiency bound calculations and are standard. Assumption CI2(c) assumes that the conditional distribution of Y given $V(X,\theta_0)$ is continuously differentiable in the index function V. This assumption combined with CI2(d) and the chain rule ensures that the conditional density of Y given $V(X,\theta)$ is differentiable in θ, so that our model is smooth, a necessary condition for efficiency calculations. Note that if Y has discrete support, this is an assumption on the smoothness of conditional probabilities rather than conditional densities.

Lemma 1.2.8 *Under* (1.2.7) *the efficient score of the model identified by* $Y \perp X | V(X, \theta_0)$ *is:*

$$S^{eff}(X, V, Y, \theta_0) = (V_\theta(X, \theta_0) - E[V_\theta(X, \theta_0)|V]) \frac{\frac{\partial}{\partial V} f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)};$$

The efficient influence function is:

$$\psi^{eff}(X, V, Y, \theta_0) = \left(V^{eff}(\theta_0)\right)^{-1} S^{eff}(X, V, Y, \theta_0)$$

where $V^{eff}(\theta_0)$*, the semiparametric efficiency bound, is:*

$$V^{eff}(\theta_0) = E\left[Var[V_\theta(X, \theta_0)|V] \left(\frac{\frac{\partial}{\partial V} f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)} \right)^2 \right]^{-1}.$$

Discussion of Lemma and Binary Choice Example

To illustrate this result, consider the binary choice model in example (1.1.5) with $Y = \mathbf{1}(\epsilon \leq X'\theta_0)$ and $\epsilon \perp X | X'\theta_0$, which implies that $Y \perp X | X'\theta_0$. Here, $V(X, \theta_0) = X'\theta_0$, and one of the elements of θ_0 is normalized to unity. The smoothness assumption here is equivalent to assuming that at ϵ continuously distributed, a frequent assumption in the identification of discrete choice models. Assumption CI2(e) also requires a finite variance for the non-normalized element of X. Using the formula above, the efficiency bound is:

$$V^{\text{eff}}(\theta_0)^{-1} = E\left[\frac{\left(\frac{\partial}{\partial \theta} P(Y = 1|X'\theta)|_{\theta=\theta_0} \right)^2}{P(Y = 1|X'\theta_0) P(Y = 0|X'\theta_0)} \right]$$

as in Cosslett (1987) and Klein and Spady (1993). This efficiency bound is the same as that for the restriction $\epsilon \perp X$, a stronger assumption which implies $Y \perp X | X'\theta_0$, as shown in Klein and Spady (1993). Since Y is a binary random variable, the conditional independence restriction is equivalent to a mean-independence or single index restriction, as in Ichimura (1993):

$$E[Y|X] = P(Y = 1|X) = P(Y = 1|X, X'\theta_0) = P(Y = 1|X'\theta_0).$$

Therefore, the efficiency bound for the conditional independence restriction is the same as the efficiency bound for the single-index restriction when Y is binary, and we see that conditional independence restrictions are a generalization of single-index restrictions. We can also consider multiple choice models with $K > 2$ alternatives where $\mathbf{Y}$ is a K by 1 vector of binary random variables such that $\mathbf{Y}_k = 1$ indicates that alternative k was selected. In this case, a conditional independence restriction of the type $\mathbf{Y} \perp X|X'\theta_0$ where X is a K by r matrix of regressors, and θ_0 is a r by 1 vector of coefficients will be sufficient for identification of θ_0 up to scale. Again, this model is equivalent to the mean-independence/single-index restriction:

$$E[\mathbf{Y}|X] = E[\mathbf{Y}|X'\theta_0].$$

As shown by Thompson (1993), the efficiency bound for using the conditional independence restriction and the stronger $\epsilon \perp X$ are different, as opposed to the binary response model. See Lee (1995) and Ruud (2000) for further details. Other semiparametric discrete choice models will be identified using conditional independence restrictions, but not under mean-independence restrictions. An example of such a model is a semi-parametric ordered choice model such as the following:

$$Y = \begin{cases} 0 \text{ if } X'\theta_0 + \epsilon < 0 \\ 1 \text{ if } X'\theta_0 + \epsilon \in [0, \mu_1) \\ 2 \text{ if } X'\theta_0 + \epsilon \in [\mu_1, \mu_2) \\ \vdots \\ J \text{ if } X'\theta_0 + \epsilon > \mu_{J-1} \end{cases}.$$

Since Y is not binary, mean-independence and conditional independence differ. It will be possible to identify θ_0 with the restriction $Y \perp X|X'\theta_0$, which is equivalent to $P(Y = j|X) = P(Y = j|X'\theta_0)$ for $j \in \{0, 1, \ldots, J\}$. More generally, we can consider the generalized regression model of Han (1987) which lets $Y = D(F(X'\theta_0, \epsilon))$, where $F(\cdot)$ is a known strictly monotone function, and $D(\cdot)$ is a known weakly monotone function. This general model includes linear regression, censored regression, duration models and discrete choice models as special cases. Letting ϵ follow a distributional single-index restriction as in Lee (1995), that is $\epsilon \perp X|X'\theta_0$, we can show that $Y \perp X|X'\theta_0$, a conditional independence restriction on

observable variables which can identify θ_0 under mild additional conditions.

1.2.4 Independence with Nuisance Function

We now consider model (1.4), which includes an unknown function $F(\cdot)$ in the residual function. We let $\rho_F(Y, W, \theta_0, F_0(\cdot))[w]$ denote the directional derivative of ρ with respect to $F(\cdot)$ in the w direction evaluated at $F_0(\cdot)$. Let $\alpha = (\theta, F(\cdot))$ and $\alpha_0 = (\theta_0, F_0(\cdot))$. To perform efficiency bound calculations, we make the following assumptions:

Assumption 1.2.9 (UI2)

(a) *(Parameter Space)* $d_\theta = 1$ *and* Θ *is compact and* $F_0 \in \mathcal{F}$*, a convex set of functions;*

(b) *(Identification)* $\rho(Y, W, \theta, F(\cdot))\perp X \Rightarrow (\theta, F(\cdot)) = (\theta_0, F_0(\cdot))$*;*

(c) *(Invertibility)* $d_Y = d_\epsilon = 1$ *and* $\rho(\cdot, W, \theta, F(\cdot))$ *is invertible for all* $(\theta, F(\cdot)) \in \Theta \times \mathcal{F}$ *a.s. - W;*

(d) *(Finite Fisher Information)* ϵ *has a continuously differentiable density function* $f_\epsilon(\cdot)$ *and* $0 < E\left[\dfrac{f'_\epsilon(\epsilon)^2}{f_\epsilon(\epsilon)^2}\right] < \infty$*;*

(e) *(Differentiability)* $\rho(Y, W, \theta, F(\cdot))$ *is differentiable with respect to* θ *a.s. -* (Y, W) *and directionally differentiable with respect to* $F(\cdot)$ *a.s. -* (Y, W)*;*

(f) *(Finite Second Moments of Derivatives)* $E\left[\|\rho_\theta(Y, W, \theta_0, F_0(\cdot))\|^2\right] < \infty$ *and* $E\left[\|\rho_F(Y, W, \theta_0, F_0(\cdot))[w]\|^2\right] < \infty$ *for all* $w \in \mathcal{F}$*.*

The assumptions presented above are similar to the assumptions in (1.2.1), but include restrictions about directional derivatives with respect to the non-parametric component. We assume in UI2(e) that the directional derivative exists, and in UI2(f) that it has finite second moment. These assumptions are equivalent to those in (1.2.1) when there is no unknown function F_0.

Lemma 1.2.10 *Under* (1.2.9) *the efficient score of the model identified by* $\rho(Y, W, \theta, F(\cdot))\perp X \Rightarrow (\theta, F(\cdot)) = (\theta_0, F_0(\cdot))$ *is:*

$$S^{eff}(X, \epsilon, \alpha_0) = E[S_\theta + S_F[w^*] + J[w^*]|X, \epsilon] - E[S_\theta + S_F[w^*] + J[w^*]|\epsilon]$$

where

$$S_\theta = S_\theta(W, \epsilon, X, \alpha_0) = \frac{\dfrac{\partial}{\partial \epsilon} f_{\epsilon,W,X}(\epsilon, W, X)}{f_{\epsilon,W,X}(\epsilon, W, X)} \rho_\theta(Y, W, \theta_0, F_0(\cdot)),$$

$$J[w] = \frac{\partial}{\partial \theta} \log \left| \frac{\partial}{\partial y'} \rho(Y, W, \alpha)|_{\alpha=\alpha_0} \right| + \log \left| \frac{\partial}{\partial y'} \rho(Y, W, \alpha_0) \right| [w],$$

$$S_F[w] = \frac{\frac{\partial}{\partial \epsilon} f_{\epsilon,W,X}(\epsilon, W, X)}{f_{\epsilon,W,X}(\epsilon, W, X)} \rho(Y, W, \theta_0, F_0(\cdot))[w],$$

and $w^* \in argmin_{w \in \mathcal{F}} E[(E[S_\theta + S_F[w] + J[w]|X, \epsilon] - E[S_\theta + S_F[w] + J[w]|\epsilon])^2]$. *The efficient influence function is:*

$$\psi^{eff}(X, \epsilon, \alpha_0) = \left(V^{eff}(\alpha_0)\right)^{-1} S^{eff}(X, \epsilon, \alpha_0)$$

where $V^{eff}(\alpha_0)$*, the semiparametric efficiency bound, is:*

$$V^{eff}(\alpha_0) = E\left[(E[S_\theta + S_F[w^*] + J[w^*]|X, \epsilon] + E[S_\theta + S_F[w^*] + J[w^*]|\epsilon])^2\right]^{-1}.$$

Discussion of Lemma and transformation IV example

We can give an interpretation to some of the terms in the efficiency bound, for example S_θ is the parametric score, $S_F[w]$ is the non-parametric score (or the score associated with the unknown function), and $J[w]$ is the jacobian term. The direction w^* can be interpreted as the "least-favorable direction", since it is the direction in which the directional derivative of the model minimizes the additional information, as can be seen from the definition of w^* as a minimizer. Note that w^* always exists since the function it minimizes is convex.

To illustrate this result, consider example (1.1.7): $Y = \Lambda_0(W\theta_0 + \epsilon)$ with $\epsilon \perp X$, where both $\Lambda_0(\cdot)$ and θ_0 are unknown. Let $\Lambda_0(\cdot)$ be a smooth and strictly increasing function from $\mathbb{R}$ to $(0, 1)$, and let $F_0(\cdot)$ denote its inverse. We assume that $F(\cdot) \in \mathcal{F}$, where $\mathcal{F}$ is the class of strictly increasing, and continuously differentiable functions from $(0, 1)$ into $\mathbb{R}$. Since $\rho(Y, W, \alpha) = F(Y) - W\theta$, this model will satisfy assumptions (1.2.9) if we assume that $E[F(Y)^2]$ for all $F \in \mathcal{F}$, ϵ has finite fisher information and W has finite variance. Mild additional regularity conditions can allow this model to be identified by the work of Torgovitsky (2012).[9]

This non-parametric transformation IV model can be motivated by a simple BLP (Berry et al. (1995)) model. To put the model in context, we let $Y \in (0, 1)$ be the market share, $W\theta_0 + \epsilon$ be the random utility level, and X be an instrument. The function $\Lambda_0(\cdot)$ is an invertible but unknown market share function that maps random utilities into market share. Many BLP models impose parametric assumptions on idiosyncratic utility shocks. They assume the shocks follow a type 1 extreme value distribution, which allows $\Lambda_0(\cdot)$ to be a logistic CDF. Recent work on these models have relaxed parametric assumptions on $\Lambda_0(\cdot)$, and have looked at the semiparametric estimation of θ_0. See Berry et al. (2012) for theoretical

[9]It is worth noting that $F_0(\cdot)$ and θ_0 are not jointly identified unless we impose additional scale and location constraints on $F_0(\cdot)$. We therefore assume that $F(1/2) = 0$ and $F'(1/2) = 1$ for all $f \in \mathcal{F}$ without loss of generality.

results on market share inversion, and Komunjer and Santos (2010) for estimation results in another simple BLP model.

To compute the efficiency bound, we use the calculations provided in the proof of Lemma (1.2.10), and we have that the parametric score is $S_\theta = -W$, the nonparametric score evaluated at direction $[w]$ is $w(Y)$, the Jacobian term is $w'(Y)$ and the least favorable direction w^* is defined implicitly as in lemma (1.2.10). The efficiency bound is then:

$$V^{\text{eff}}(\alpha_0) = E\left[(E[S_\theta + S_F[w^*] + J[w^*]|X,\epsilon] - E[S_\theta + S_F[w^*] + J[w^*]|\epsilon])^2\right]^{-1},$$

where

$$\begin{aligned} &E[S_\theta + S_F[w] + J[w]|X,\epsilon] - E[S_\theta + S_F[w] + J[w]|\epsilon] = \\ &\qquad (E[W - w(Y)|X,\epsilon] - E[W - w(Y)|\epsilon])\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)} \\ &+\frac{\partial}{\partial\epsilon}(E[W - w(Y)|X,\epsilon] - E[W - w(Y)|\epsilon]) + E[w'(Y)|X,\epsilon] - E[w'(Y)|\epsilon]. \end{aligned}$$

Note that if there is no unknown function in $\rho(\cdot)$, meaning that $F_0(\cdot)$ is assumed known, all terms with $[w]$ disappear, and the efficiency bound becomes that of the linear IV model. In that case, the bound will depend solely on the joint distribution of (W, X, ϵ), and on the value θ_0, while the known transformation $F_0(\cdot)$ will not affect the bound. Since the efficiency bound depends on whether or not $F_0(\cdot)$ is known, it will not be possible to estimate θ_0 adaptively. To proceed with the estimation of θ_0 in this example, one must consider $F_0(\cdot)$ as an infinite dimensional nuisance function that must be estimated alongside θ_0. Also note that w^* often does not have a closed form expression, as it is defined as a function that minimizes a criterion function.

1.3 Estimation

The estimator for model (1.2) is based on a criterion function containing a number of covariances between basis functions of X and $\rho(Y, W, \theta_0)$. We define the mean-zero vector of basis functions of X as $\hat{q}^K(X_i) = \hat{q}^K_i$, a $K \times 1$ vector of the form $\hat{q}^K_i = r^K_i - \frac{1}{N}\sum_{j=1}^N r^K_j$, with r^K_j a $K \times 1$ vector of basis functions of X. Let $q^K_i = r^K_i - E[r^K_i]$. Let $\hat{\rho}^L(Y_i, W_i, \theta) = \hat{\rho}^L_i(\theta) = p^L_i(\theta) - \frac{1}{N}\sum_{j=1}^N p^L_j(\theta)$ where $p^L_i(\theta)$ is a $L \times 1$ vector of basis functions of $\rho(Y_i, W_i, \theta)$. For example, they could be the first L powers of $\rho(Y_i, W_i, \theta)$, stacked in a vector. Also, let $\rho^L_i(\theta) = p^L_i(\theta) - E[p^L_j(\theta)]$.

Estimation is based on the following zero-covariance conditions:

$$
\begin{aligned}
E[p^L(Y_i, W_i, \theta) \otimes r^K(X_i)] - E[p^L(Y_i, W_i, \theta)] \otimes E[r^K(X_i)] &= 0 \\
\Rightarrow E[\rho^L(Y_i, W_i, \theta) \otimes q^K(X_i)] &= 0 \\
\Rightarrow E[g_i^{KL}(\theta)] &= 0
\end{aligned}
$$

a $KL \times 1$ vector of moment conditions defined by $g_i^{KL}(\theta) = \rho^L(Y_i, W_i, \theta) \otimes q^K(X_i)$. Note that this is not exactly a GMM problem: the function g_i^{KL} is not known a priori, since both $q^K(X_i)$ and $\rho^L(Y_i, W_i, \theta)$ involve expectations that need to be estimated. Essentially, this problem relies on covariance conditions rather than on moment conditions. We will show that the GMM framework can also be used to deliver asymptotically efficient estimates of parameters identified through covariance conditions by modifying the optimal weighting matrix to reflect the covariance structure of the estimating equations. Assuming K and L fixed, the optimal GMM estimator of θ_0 based on these KL moment conditions is:

$$
\begin{aligned}
\hat{\theta} &= \text{argmin}_{\theta \in \Theta} \hat{R}^{KL}(\theta) \\
\hat{R}^{KL}(\theta) &= \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i^{KL}(\theta)' \left(\frac{1}{N} \sum_{i=1}^{N} [\hat{g}_i^{KL}(\tilde{\theta}) \hat{g}_i^{KL}(\tilde{\theta})'] \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i^{KL}(\theta) \\
&= \hat{\bar{g}}^{KL}(\theta)' \hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1} \hat{\bar{g}}^{KL}(\theta)
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{g}_i^{KL}(\theta) &= \hat{\rho}^L(Y_i, W_i, \theta) \otimes \hat{q}^K(X_i) \\
\hat{\bar{g}}^{KL}(\theta) &= \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i^{KL}(\theta) \\
\hat{\bar{\Omega}}^{KL}(\theta) &= \frac{1}{N} \sum_{i=1}^{N} [\hat{g}_i^{KL}(\theta) \hat{g}_i^{KL}(\theta)'] \\
\bar{\Omega}^{KL}(\theta) &= \frac{1}{N} \sum_{i=1}^{N} [g_i^{KL}(\theta) g_i^{KL}(\theta)'] \\
\tilde{\theta} &= \theta_0 + O_p(\tau_N) \text{ with } \tau_N \to 0
\end{aligned}
$$

Here, $\tilde{\theta}$ is a preliminary "first-step" estimator of θ_0 that is usually, but not necessarily, $\sqrt{N}$ consistent, and $\hat{\bar{\Omega}}^{KL}(\tilde{\theta})$ is the estimate of the variance-covariance matrix of the sample

covariances. To ensure identification and consistency, we need to make further assumptions about the basis functions:

Assumption 1.3.1 (Basis Functions)

(a) *(X Basis Functions) $d_X = 1$, $\sup_{x \in X} \|r^K(x)\| \leq \zeta(K)$ with $\sqrt{K} \leq \zeta(K)$, and $E[r_i^K r_i^{K\prime}]$ has its smallest eigenvalue uniformly bounded away from 0;*

(b) *($\rho(Y, W, \theta)$ Basis Functions) $\sup_{(y,w) \in \mathcal{Y} \times \mathcal{W}} \|p^L(\theta_0)(y, w)\| \leq \zeta(L)$ with $\sqrt{L} \leq \zeta(L)$, and $E[p_i^L(\theta) p_i^L(\theta)']$ has its smallest eigenvalue uniformly bounded away from 0 for any θ in a neighborhood of θ_0;*

(c) *(Spanning) For each δ and continuous function $v(x, \rho(y, w, \theta)) \in \mathbb{R}$, there exists K, L such that $\|v(x, \rho(y, w, \theta)) - A'(r^K(x) \otimes \rho^L(y, w, \theta))\| < \delta$ for a $KL \times 1$ vector A.*

Assumptions (1.3.1)(a) and (b) are rate conditions on the approximation functions. When using complex exponentials on a bounded domain as basis functions, $\zeta(K) = C\sqrt{K}$ for a bounded constant C, and the eigenvalue conditions are satisfied. If we were to use power functions instead, $\zeta(K)$ would be directly proportional to K and restrictions on the range of X and $\rho(Y, W, \theta)$ are needed as well. See Andrews (1991), Gallant and Souza (1991), Newey (1997) and Donald et al. (2003) for more details. Without loss of generality we can assume that both $\rho(Y, W, \theta)$ and X are bounded: independence of two random variables X and Y is equivalent to independence of $\Phi(X)$ and $\Phi(Y)$, where Φ is a bounded, invertible function mapping from $\mathbb{R}$ to a bounded interval (e.g. the CDF of a continuously distributed random variable). Assumption (1.3.1)(c) requires the basis functions to be complete, meaning that they approximate arbitrarily well functions of their arguments.

We find an alternative representation for the independence restriction in model (1.2), using the following definition:

Definition 1.3.2 (Measure-Determining Class) *Let $\mathcal{F}$ be a class of functions, and X be a random variable. If there exists a bijection between $\{E[f(X)] : f \in \mathcal{F}\}$ and the distribution function of X, F_X, we say that $\mathcal{F}$ is a measure determining class for X.*

Well known examples measure-determining classes include $\{e^{it'X} : t \in \mathbb{R}^{d_X}\}$, $\{\prod_{l=1}^{d_X} \mathbf{1}(X_i \leq t_l) : t \in \mathbb{R}^{d_X}\}$, and the set of all continuous bounded functions of X. If X has discrete support, the set of indicators for all support points (minus one) of X is also a convergence determining class. Independence of two random variables can be characterized as an uncorrelatedness condition between functions in measure-determining classes:

Lemma 1.3.3 *Let $X \in \mathbb{R}^{d_X}$ and $Y \in \mathbb{R}^{d_Y}$ be random variables. Let $\mathcal{F}$ and $\mathcal{G}$ be measure-determining classes of dimension d_X and d_Y respectively. Then $Cov(f(X), g(Y)) = 0$ for all $(f, g) \in \mathcal{F} \times \mathcal{G}$ if and only if $X \perp Y$.*

Therefore, going back to the setup in (1.2), we restate the independence restriction as the following:

$$\text{Cov}\left(f(\rho(Y,W,\theta_0)), g(X)\right) = 0 \text{ for all } (f,g) \in \mathcal{F} \times \mathcal{G}. \tag{1.6}$$

In the remainder of this chapter, we will focus on $\mathcal{F}_0^{d_\epsilon} = \{e^{is'\cdot} : s \in \mathbb{R}^{d_\epsilon}\}$ and $\mathcal{F}_0^{d_X} = \{e^{it'\cdot} : t \in \mathbb{R}^{d_X}\}$ when they are continuous random variables, and on indicators at support points if they are discrete random variables. Unlike moment-generating functions, equality of characteristic functions on a positive measure interval does not imply equality everywhere on the Euclidean space: one can construct two characteristic functions that are equal on arbitrarily large intervals, but different elsewhere.[10] Also, characteristic functions exist and are bounded for all random variables.

We will assume that we are working with Fourier series basis functions, and that $r_i^K = e^{i\vec{s}_K X_i}$, where $\vec{s}_K = [s_1, \ldots, s_K]'$ is a $K \times 1$ vector such that $\vec{s}_{K+1} = [\vec{s}_K', s_{K+1}]'$ and as $K \to \infty$, $\vec{s}_K$ is dense in $\mathbb{R}$.[11] Similarly, $p_i^L(\theta) = e^{i\vec{t}_L \rho(Y_i, W_i, \theta)}$, and $\vec{t}_L$ is a $L \times 1$ vector that becomes dense in $\mathbb{R}$ as $L \to \infty$. This basis will satisfy assumptions (1.3.1)(a)-(b) with $\zeta(K) = O(\sqrt{K})$ and $\zeta(L) = O(\sqrt{L})$. The asymptotic denseness of the vectors will allow assumption (1.3.1)(c) to be satisfied. The choice of basis and its implications are beyond the scope of this chapter.

1.3.1 Consistency and Asymptotic Normality

Before deriving the estimator's asymptotic properties, we make more assumptions about regularity conditions. Denote by $\rho_\theta^L(\theta)$ the $L \times 1$ vector of derivatives of $\rho^L(\theta)$ with respect to θ.

Assumption 1.3.4 (Regularity Conditions)

(a) *(Identification)* $\rho(Y,W,\theta) \perp X \implies \theta = \theta_0$*;*

(b) *(IID)* $\{X_i, Y_i, W_i\}_{i=1}^N$ *are identically and independently distributed;*

(c) *(Compact Parameter Space)* $\theta \in Int(\Theta)$ *and* Θ *is compact;*

(d) *(Differentiability and Global Lipschitz of Residual Function)* $\rho(Y,W,\theta)$ *is twice differentiable in a neighborhood of* θ_0 *and* $\|\rho^L(Y,W,\tilde{\theta}) - \rho^L(Y,W,\theta)\| \leq \delta_L(Y,W)|\tilde{\theta} - \theta|$ *and* $E[\delta_L(Y,W)^2|X] = O_p(L)$ *for any* $\tilde{\theta}, \theta$ *in* Θ*. Also,* $\|\rho_\theta^L(Y,W,\tilde{\theta}) - \rho_\theta^L(Y,W,\theta)\| \leq \delta_{\theta L}(Y,W)|\tilde{\theta} - \theta|$ *and* $E[\delta_{\theta L}(Y,W)^2|X] = O_p(L)$ *for any* $\tilde{\theta}, \theta$ *in* Θ*.*

(e) *(Invertibility)* $\rho(\cdot, W, \theta)$ *is differentiable and invertible for all* $\theta \in \Theta$*, a.s. -* W*, and its inverse function* $Y = m(\cdot, W, \theta)$ *is continuous in* θ*;*

[10] If we restrict our attention to random variables with analytic characteristic functions, equality on an interval with positive Lebesgue measure will imply equality everywhere. Examples of distributions with analytic characteristic functions include the normal and Laplace distributions. See Lukacs (1960).

[11] Throughout the chapter, we will use the notation $e^{\vec{a}}$ to denote the element by element exponentiation of the vector $\vec{a}$.

(f) *(Preliminary Estimator)* $\tilde{\theta}_N$ *is a preliminary and consistent estimator of* θ_0 *satisfying* $\|\tilde{\theta}_N - \theta_0\| = O_p(\tau_N)$*, with* $\tau_N \to 0$ *as* $N \to \infty$*;*

(g) *(Finite Second Moments)* ϵ *has a continuously differentiable density function* $f_\epsilon(\cdot)$ *and* $0 < E\left[\frac{f'_\epsilon(\epsilon)^2}{f_\epsilon(\epsilon)^2}\right] < \infty$*. Also,* $E[\|\rho_\theta(Y, W, \theta_0)\|^2] < \infty$*;*

(h) *(Global Lipschitz Objective Function)* $\sup_{\theta,\tilde{\theta}\in\Theta} |\hat{R}^{KL}(\theta) - \hat{R}^{KL}(\tilde{\theta})| \leq \hat{D}|\theta - \tilde{\theta}|^\alpha$*, where* $\hat{D} = O_p(1)$ *and* $\alpha > 0$*.*

Most of these assumptions are similar to those in (1.2.1). Assumption (1.3.4)(d) is strong, but is satisfied when $\rho(Y, W, \theta)$ is twice differentiable and additional mild conditions are satisfied when using Fourier series. In (1.3.4)(f), we require the existence of a preliminary consistent estimator. In the unconditional independence case, most parameters are overidentified, therefore using an estimate coming from a fixed and finite number of moment conditions will usually yield a consistent and $\sqrt{N}$ consistent estimate. Assumption (1.3.4)(h) is strong, and implies stochastic equicontinuity of the objective function. We are making progress in trying to relax this condition with the help of lower-level assumptions. Similar assumptions to those above can be found in Donald et al. (2003). Asymptotic normality and consistency require that both K and L increase towards infinity along with the sample size, but they need to increase at a rate slow enough that the asymptotic bias of $\widehat{\theta}$ coming from the increasing number of moment conditions goes to 0.

Theorem 1.3.5 (Consistency and Asymptotic Normality) *Let Assumptions* (1.3.1) *and* (1.3.4) *hold,* $K^2L^2(\tau_N + \frac{1}{\sqrt{N}}) \to 0$*, and* $K, L \to \infty$ *as* $N \to \infty$*, then* $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ *and* $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V_{\mathit{eff}}(\theta_0))$*, where* $V_{\mathit{eff}}(\theta_0)$ *is defined in Lemma* (1.2.2)*.*

A proof of this result is given in the appendix. The proof follows the methods in Newey and McFadden (1994)'s chapter. Furthermore, we can derive a consistent optimal variance estimate as follows:

$$\widehat{V} = \left[\left(\frac{\partial}{\partial\theta}\hat{g}_i^{KL}\left(\widehat{\theta}\right)\right)'\left(\widehat{\Omega}^{KL}\left(\tilde{\theta}\right)\right)^{-1}\left(\frac{\partial}{\partial\theta}\hat{g}_i^{KL}\left(\widehat{\theta}\right)\right)\right]^{-1}.$$

Heuristically, this variance estimator will converge if the same rate conditions are satisfied.

These asymptotic results do not provide guidance as to the exact choice of rate for both K and L since they only provide maximal rates of growth. For comparison's sake, the GMM-type estimator for mean-independence restrictions in Donald et al. (2003) required that $\frac{K\zeta(K)^2}{N} \to 0$, which will be equivalent to $K^2/N \to 0$, while we require $K^2L^2/\sqrt{N} \to 0$, so if we let $K = L$, K must be of order $o\left(N^{\frac{1}{8}}\right)$, while in Donald et al. (2003) K must satisfy $K = o\left(N^{\frac{1}{2}}\right)$.

To prove efficiency of this estimator using zero-covariance restrictions, we recast covariance conditions as moment conditions by introducing additional ancillary parameters. We then use the efficiency properties of GMM under moment conditions. For example $\text{Cov}\left(h(X,\theta),Y\right) = E[h(X,\theta)(Y - E[Y])] = 0$ can be recasted as two moment conditions with two parameters, the new parameter being $E[Y]$. This also illustrates an alternative method for computing the efficiency bound of models with independence restrictions which involves the computation of the limit of the efficient GMM variance for fixed K and L.

When estimating models based on independence restrictions, a commonly used objective function is the Cramer von-Mises distance between the joint distribution of $\rho(Y,W,\theta)$ and X, and the product of their marginal distributions:

$$\begin{aligned}\hat{\theta}_{\text{VM}} &= \text{argmin}_{\theta\in\Theta} \int\int (\hat{F}_{\rho(Y,W,\theta),X}(s,t) - \hat{F}_{\rho(Y,W,\theta)}(s)\hat{F}_X(t))^2 d\mu(s,t) \\ &= \text{argmin}_{\theta\in\Theta} \int\int (\widehat{\text{Cov}}(1(\rho(Y,W,\theta)\leq s), 1(X\leq t)))^2 d\mu(s,t)\end{aligned}$$

where $\mu(s,t)$ is a probability distribution on $\mathbb{R}^2$ specified by the econometrician. See Manski (1983), Brown and Wegkamp (2002), Domínguez and Lobato (2004) and Komunjer and Santos (2010). Since this integral cannot be evaluated directly, it can be approximated through a discrete approximation:

$$\hat{\theta}^K_{\text{VM}} = \text{argmin}_{\theta\in\Theta} h^K(\theta)' W^K h^K(\theta)$$

where $h^k(\theta) = [\widehat{\text{Cov}}(1(\rho(Y,W,\theta)\leq s_1), 1(X\leq t_1)), \widehat{\text{Cov}}(1(\rho(Y,W,\theta)\leq s_2), 1(X\leq t_1))\ldots]'$ is a $K^2\times 1$ vector, and $W^K = \text{diag}\{\mu(s_1,t_1),\mu(s_2,t_1),\ldots\mu(s_K,t_K)\}$. Therefore, CV-M estimation of θ_0 is similar to our approach, except that it requires a diagonal weighting matrix. Since the optimal weighting matrix is usually non-diagonal, CV-M estimation cannot reach the efficiency bound, even through a judicious choice for $\mu(\cdot,\cdot)$ and large K.

1.3.2 Example and Discussion

Going back to the linear regression example (1.1.1), the efficient estimator we propose would use the following covariances as a basis for estimation:

$$E[e^{i\vec{t}_L(Y_i - X_i\theta)}\otimes e^{i\vec{s}_K X_i}] - E[e^{i\vec{t}_L(Y_i - X_i\theta)}]\otimes E[e^{i\vec{s}_K X_i}] = 0.$$

These KL covariances form the basis of estimation, and the user must select K, L and the

vectors $\vec{s}_K$ and $\vec{t}_L$. K and L must satisfy the rate requirements specified in the consistency and asymptotic normality theorem, and $\vec{s}_K$ and $\vec{t}_L$ are only constrained by the denseness condition. For fixed K and L, the variance of the GMM estimator is:

$$V_{KL}^{-1}(\theta_0) = \text{Cov}\left(\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}, e^{i\vec{t}_L\epsilon}\right)' \text{Var}\left[e^{i\vec{t}_L\epsilon}\right]^{-1} \text{Cov}\left(\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}, e^{i\vec{t}_L\epsilon}\right) \times$$
$$\text{Cov}\left(X, e^{i\vec{s}_K X}\right)' \text{Var}\left[e^{i\vec{s}_K X}\right]^{-1} \text{Cov}\left(X, e^{i\vec{s}_K X}\right)$$

From this expression, we can see that $\text{Cov}\left(\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}, e^{i\vec{t}_L\epsilon}\right)' \text{Var}\left[e^{i\vec{t}_L\epsilon}\right]^{-1} \text{Cov}\left(\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}, e^{i\vec{t}_L\epsilon}\right)$ is the variance of the projection of $\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ on L complex exponential functions. We checked earlier that complex exponentials -or equivalently sines and cosines- can approximate continuous functions arbitrarily well in mean-square. This directly translates into the variance of the projection of $\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ converging to the variance of $\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$ itself, as $L \to \infty$. A similar argument will yield that $\text{Cov}\left(X, e^{i\vec{s}_K X}\right)' \text{Var}\left[e^{i\vec{s}_K X}\right]^{-1} \text{Cov}\left(X, e^{i\vec{s}_K X}\right)$ converges to the variance of X as $K \to \infty$. So we see that even if K and L are not allowed to grow with the sample size, $V_{KL}(\theta_0) \to V^{\text{eff}}(\theta_0)$. We also get an ϵ-efficiency result as in Chamberlain (1992), that is for every $\epsilon > 0$, there exists K and L large enough so that

$$\|V_{KL}(\theta_0) - V^{\text{eff}}(\theta_0)\| < \epsilon.$$

Reasons to let K and L go to infinity are twofold. The first is to ensure that the estimator's asymptotic variance attains the efficiency bound asymptotically. The second is that by letting K and L go to infinity and in turn letting the vectors $\vec{s}_K$ and $\vec{t}_L$ span the real line, we are insuring the full use of the independence restriction. Full independence can potentially be a necessary condition for the identification of θ_0, since an arbitrary set of covariance conditions might not provide global identification. Domínguez and Lobato (2004) explore this question with respect to mean-independence restrictions.

A useful feature of relying on a GMM framework for estimation is that we can append additional moment conditions efficiently, and derive the efficiency bound without relying on the projection method. One potential application of this feature is to the model of Brown and Newey (1998). They focus on the efficiency bound of $\theta_0 = E[m(X, \beta_0)]$ where $m(\cdot)$ is a known function, and β_0 is identified through $\rho(X, \beta) \perp X$. We can consider the following model, a slight generalization of their setup:

$$\left(\begin{array}{c} \rho(Y,W,\beta) \perp X \\ E[m(X,Y,Z,W,\beta) - \theta] = 0 \end{array} \right) \Longrightarrow (\beta,\theta) = (\beta_0,\theta_0).$$

Using the framework in the previous sections, we can convert this independence restriction in an increasing number of covariance conditions, and append the extra moment condition that identifies θ_0 to it. Setting up a system of covariance and moment restrictions, we can recover the efficiency bound derived by Brown and Newey (1998). Their approach consists of approximating the efficient score using a series estimate, and then using a V-statistic type criterion function. It is simple in the GMM framework to efficiently add moment conditions, and possibly multiple independence, mean-independence or covariance restrictions in a single system.

1.3.3 Feasible GMM Estimation Under Conditional Independence Restrictions

We generalize our setup to include models that are identified through conditional independence restrictions, such as model (1.1). One can do this similarly to the unconditional independence case, by using the restriction:

$$\begin{aligned} & \rho(Y,W,\theta) \perp X|Z \\ \Longleftrightarrow & \text{Cov}\left(e^{it\rho(Y,W,\theta)}, e^{isX}|Z\right) = 0 \\ \Longleftrightarrow & E[e^{it\rho(Y,W,\theta)}(e^{isX} - E[e^{isX}|Z])|Z] = 0 \\ \Longleftrightarrow & E[e^{it\rho(Y,W,\theta)}(e^{isX} - E[e^{isX}|Z])e^{iuZ}] = 0 \end{aligned}$$

for all $(s,t,u) \in \mathbb{R}^3$.

Using this method requires the computation of a preliminary non-parametric estimator of $E[e^{isX}|Z]$, which will not affect the efficiency bound if appropriate convergence conditions are imposed. In many semi-parametric problems, we need the estimator to be $o_p(N^{-1/4})$. See Robinson (1988) for an example. The feasible estimator of θ based on this equivalence will be based on:

$$E[e^{i\vec{t}_L\rho(Y,W,\theta)} \otimes (e^{i\vec{s}_K X} - \hat{\lambda}(\vec{s}_K, Z)) \otimes e^{i\vec{u}_M Z}] = 0$$

where $\hat{\lambda}(\vec{s}_K, Z)$ is a preliminary estimator of $\lambda(\vec{s}_K, Z) = E[e^{i\vec{s}_K X}|Z]$. When Z is discrete, we do not need to use a preliminary estimator, and can add some moment conditions to jointly

estimate this conditional expectation with the other moments. For example, if Z is a binary variable taking the values of 0 or 1, we can replace the term $e^{i\vec{u}_M Z}$ by an indicator for Z being equal to 1, since the distribution of this indicator is in a bijection with the distribution of Z.[12] When Z's distribution is continuous, we need to use an approximating method such as kernels, series or splines to estimate $\lambda(\vec{s}_K, Z)$ in a preliminary step. In keeping with the spirit of the second step, it is possible to estimate $\lambda(\vec{s}_K, Z)$ using a series estimator with complex exponential terms.

Estimator

We must select K, L and M, the dimensions of the vectors $\vec{s}_K, \vec{t}_L$ and $\vec{u}_M$ respectively. Also, we let $\hat{\lambda}(\vec{s}_K, Z)$ be the preliminary estimator of $E[e^{i\vec{s}_K}|Z]$. We must then compute the asymptotic variance $V(\theta_0)$ of the moment conditions:

$$\hat{h}^{KLM}(\theta_0) = \frac{1}{N}\sum_{i=1}^{N}[e^{i\vec{t}_L\rho(Y_i,W_i,\theta_0)} \otimes (e^{i\vec{s}_K X_i} - \widehat{\lambda}(\vec{s}_K, Z_i)) \otimes e^{i\vec{u}_M Z_i}]$$
$$\sqrt{N}\hat{h}^{KLM}(\theta_0) \xrightarrow{d} N(0, V^{KLM}(\theta_0))$$

Then the feasible estimator $\hat{\theta}$ is defined by:

$$\widehat{\theta} = \text{argmin}_\theta \widehat{R}(\theta)^{KLM}$$
$$\widehat{R}^{KLM}(\theta) = \widehat{h}^{KLM}(\theta)'(\widehat{V}^{KLM}(\tilde{\theta}))^{-1}\widehat{h}^{KLM}(\theta)$$

where $\hat{V}^{KLM}(\theta) = \frac{1}{N}\sum_{i=1}^{N}[(e^{i\vec{t}_L\rho(Y_i,W_i,\theta)} \otimes (e^{i\vec{s}_K X_i} - \widehat{\lambda}(\vec{s}_K, Z_i)) \otimes e^{i\vec{u}_M Z_i})(e^{i\vec{t}_L\rho(Y_i,W_i,\theta)} \otimes (e^{i\vec{s}_K X_i} - \widehat{\lambda}(\vec{s}_K, Z_i)) \otimes e^{i\vec{u}_M Z_i})']$ is an estimate of the matrix $V^{KLM}(\theta_0)$ defined above, and $\tilde{\theta}$ is a preliminary and consistent estimator for θ_0.

Consistency and Asymptotic Normality

To prove consistency and asymptotic normality of the estimator defined above, we assume the following:

Assumption 1.3.6 (Regularity Conditions)

(a) *(Identification)* $\rho(Y, W, \theta) \perp X|Z \Longrightarrow \theta = \theta_0$*;*

(b) *(IID)* $\{X_i, Y_i, W_i, Z_i\}_{i=1}^N$ *are identically and independently distributed;*

[12] trivially, since the indicator is equal to Z itself.

(c) *(Compact Parameter Space)* $\theta \in Int(\Theta)$ *and* Θ *is compact;*

(d) *(Differentiability and Global Lipschitz of Residual Function)* $\rho(Y,W,\theta)$ *is twice differentiable in a neighborhood of* θ_0 *and* $\|\rho^L(Y,W,\tilde{\theta}) - \rho(Y,W,\theta)\| \leq \delta(Y,W)|\tilde{\theta} - \theta|$ *and* $E[\delta(Y,W)^2|X] = O_p(L)$ *for any* $\tilde{\theta}, \theta$ *in* Θ. *Also,* $\|\rho_\theta^L(Y,W,\tilde{\theta}) - \rho_\theta(Y,W,\theta)\| \leq \delta_\theta(Y,W)|\tilde{\theta} - \theta|$ *and* $E[\delta_\theta(Y,W)^2|X] = O_p(L)$ *for any* $\tilde{\theta}, \theta$ *in* Θ.

(e) *(Invertibility)* $\rho(\cdot,W,\theta)$ *is differentiable and invertible for all* $\theta \in \Theta$, *a.s - W, and the inverse function* $Y = m(\cdot,W,\theta)$ *is continuous in* θ;

(f) *(Preliminary Estimator)* $\tilde{\theta}_N$ *is a preliminary and consistent estimator of* θ_0 *satisfying* $\|\tilde{\theta}_N - \theta_0\| = O_p(\tau_N)$, *with* $\tau_N \to 0$ *as* $N \to \infty$;

(g) *(Finite Second Moments)* ϵ *has a continuously differentiable density function* $f_\epsilon(\cdot)$ *and* $0 < E\left[\dfrac{f'_{\epsilon|Z}(\epsilon|Z)^2}{f_{\epsilon|Z}(\epsilon|Z)^2}\right] < \infty$. *Also,* $E[\|\rho_\theta(Y,W,\theta_0)\|^2] < \infty$;

(h) *(Nonparametric First Stage)* $\|\widehat{\lambda}(\vec{s}_K, Z) - \lambda(\vec{s}_K, Z)\| = O_p(\nu_N) = o_p(\sqrt{K}N^{-1/4})$.

(i) *(Global Lipschitz Objective Function)* $\sup_{\theta,\tilde{\theta}\in\Theta} |\hat{R}^{KLM}(\theta) - \hat{R}^{KLM}(\tilde{\theta})| \leq \hat{D}|\theta - \tilde{\theta}|^\alpha$, *where* $\hat{D} = O_p(1)$ *and* $\alpha > 0$.

These assumptions are similar to those in (1.3.4), and assumption (1.3.6)(h) is akin to standard regularity condition for non-parametric terms in GMM problems. Since the vector $\lambda(\vec{s}_K, Z)$ has K elements, we require that the approximation error increases with K at rate $\sqrt{K}$, but decreases with N at rate $N^{-1/4}$. These growth rate of K can be selected appropriately so that the estimation error from the non-parametric component is asymptotically negligible. Assumption (1.3.6)(i) is strong, and implies stochastic equicontinuity of the objective function in this model. We are also making progress in trying to relax this condition with the help of lower-level assumptions. We now present the asymptotic properties of the estimator.

Theorem 1.3.7 (Estimator under Conditional Independence Restrictions) *Let* (1.3.6) *hold,* $K^2L^2M^2(\tau_N + \frac{1}{\sqrt{N}}) \to 0$, *and* $K, L, M \to \infty$ *as* $N \to \infty$, *then* $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ *and* $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V^{eff}(\theta_0))$, *where* $V^{eff}(\theta_0)$ *is defined in* (1.2.4).

This theorem shows conditions under which the estimator will attain the efficiency bound. If $K = L = M$ and the preliminary estimator is $\sqrt{N}$ consistent, we must have that $K = o(N^{1/12})$ to ensure that the bias term goes away. Again, efficiency is achieved from letting K, L and M increase, since this lets us use asymptotically all the information contained in the conditional independence restriction.

1.4 Feasible GMM Estimation Under Independence Restrictions containing Unknown Functions

In this section, we conjecture a method which can be used to deliver an efficient estimator for θ_0 under $\rho(Y, W, \theta_0, F_0(\cdot)) = \epsilon \perp X$. Ai and Chen (2003) propose a method to compute efficient estimates for θ_0 under $E[\rho(Y, W, \theta_0, F_0(\cdot))|X] = 0$, with $F_0(\cdot)$ unknown. Their method relies on approximating $F_0(\cdot)$ with basis functions, so that the minimization is over a finite dimensional sieve space, rather than an infinite dimensional functional space. For exposition, we will approximate $F_0(\cdot)$ with power series, and let $F_0(\cdot)$ be a mapping between $\mathbb{R}$ and $\mathbb{R}$. Therefore, $\widehat{F}_J(a, \phi) = \phi_1 a + \phi_2 a^2 + \ldots + \phi_J a^J$, where ϕ is a J by 1 vector of coefficients which multiply the basis functions approximating $F(\cdot)$. Note that $\epsilon \perp X$ is equivalent to

$$E[e^{is\epsilon}|X] = E[e^{is\epsilon}]$$

for all $s \in \mathbb{R}$. We can make use of Ai and Chen (2003)'s framework, and consider an infinite number of conditional mean restrictions, indexed by $s \in \mathbb{R}$, such that the estimating equation is:

$$E[e^{is\rho(Y,W,\theta,F(\cdot))} - E[e^{is\rho(Y,W,\theta,F(\cdot))}]|X] = 0 \Longleftrightarrow (\theta, F(\cdot)) = (\theta_0, F_0(\cdot)).$$

To fix ideas, we will consider the model $Y = X\theta_0 + g_0(W) + \epsilon$ with $\epsilon \perp (X, W)$. Using the optimal weighting matrix provided in Ai and Chen (2003), we can see that the efficiency bound of this model will be:

$$\begin{aligned}
\Sigma_0 f(\cdot)(s) &= \int \text{Cov}\left(e^{is\epsilon}, e^{is'\epsilon}\right) f(s')ds', \\
\|f(.)\|^2_{\Sigma_0} &= \|\Sigma_0^{-1} f(\cdot)(s)\|^2, \\
D_{w0}(X, W, s) &= (-XisE[e^{is\epsilon}] + isE[e^{is\epsilon}]E[X|W]), \\
V^{\text{eff}}(\alpha_0) &= E[\|D_{w0}(X, W, s)\|^2_{\Sigma_0}]^{-1},
\end{aligned}$$

where $D_{w0}(X, W, s)$ is a random function of X and W indexed by $s \in \mathbb{R}$, Σ_0 is an operator on the set of functions $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R}\}$ with kernel $k(s, s') = \text{Cov}\left(e^{is\epsilon}, e^{is'\epsilon}\right)$, and $\|\cdot\|^2_{\Sigma_0}$ is the reproducing kernel Hilbert space norm associated with the operator Σ_0. We have shown in the appendix that $\|isE[e^{is\epsilon}]\|^2_{\Sigma_0} = E\left[\left(\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}\right)^2\right]^{-1}$, and therefore the quantity V_{eff} is equal to $\text{Var}\left[X - E[X|W]\right]^{-1} E\left[\left(\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}\right)^2\right]$. This bound is equal to the one computed using the projection method in section 2.

To make this estimator operational, we could let

$$(\widehat{\theta}, \widehat{\phi}_J) = \text{argmin}_{\theta \in \Theta, \phi_J \in \Phi_J} \widehat{E}\left[\widehat{m}(X_i)'\widehat{\Sigma}(X_i)\widehat{m}(X_i)\right],$$

where Φ_J is a sieve space that becomes dense in Φ, the function space in which $F_0(\cdot)$ resides, as $J \to \infty$. $\widehat{m}(X) = \widehat{E}[e^{i\vec{s}_K \rho(Y_i, W_i, \theta, F(\cdot))} - \widehat{E}[e^{i\vec{s}_K \rho(Y_i, W_i, \theta, F(\cdot))}]|X_i]$, and the conditional expectation is computed using either a series approach (as in Ai and Chen (2003)) or kernel methods. Also,

$$\widehat{\Sigma}(X_i) = \widehat{E}\Big[(e^{i\vec{s}_K \rho(Y_i, W_i, \tilde{\theta}, \tilde{F}(\cdot))} - \widehat{E}[e^{i\vec{s}_K \rho(Y_i, W_i, \tilde{\theta}, \tilde{F}(\cdot))}])$$
$$(e^{i\vec{s}_K \rho(Y_i, W_i, \tilde{\theta}, \tilde{F}(\cdot))} - \widehat{E}[e^{i\vec{s}_K \rho(Y_i, W_i, \tilde{\theta}, \tilde{F}(\cdot))}])' | X_i\Big],$$

where $(\tilde{\theta}, \tilde{F}(\cdot))$ is a consistent first step estimator for $(\theta_0, F_0(\cdot))$. Further work is needed to list regularity conditions sufficient for consistency and asymptotic normality of the proposed efficient estimator.

1.5 Monte-Carlo Study

To investigate the finite sample performance of the estimator, we perform Monte Carlo studies based on the unconditional independence restriction (1.2) for different data-generating processes. Our first study is based on the design of Hsieh and Manski (1987) and investigates the simple exogenous linear regression model $Y = X\beta_0 + \epsilon$ with ϵ independent from X. We let X be normally distributed with mean 0 and variance 1, and we set the parameter of interest $\beta_0 = 1$. We consider four types of distribution for the error ϵ: (a) a standard normal distribution, (b) a 50/50 mixture combining $N(-1, 1)$ and $N(1, 4)$, two independent normal distributions, (c) a mean zero Student's-t distribution with three degrees of freedom and (d), a Laplace (Double Exponential) with mean 0 and scale parameter equal to 1.

Estimation is based on the following covariance restrictions:

$$E[e^{i\vec{t}_L(Y_i - X_i\theta)} \otimes e^{i\vec{s}_K X_i}] - E[e^{i\vec{t}_L(Y_i - X_i\theta)}] \otimes E[e^{i\vec{s}_K X_i}] = 0.$$

In practice, K and L must be selected, as well as $\vec{s}_K$ and $\vec{t}_L$. We consider estimation with $K = L$, and we let them range from 1 to 6, meaning that between 1 and 36 covariance conditions are used for the estimation of θ_0. For a fixed K, we let $\vec{s}_K$ be the inverse standard normal CDF evaluated at K equally spaced points on the $[0, 1]$ interval.[13] $\vec{s}_K$ and $\vec{t}_L$ are theoretically unrestricted, except by the condition that requires them to become dense in $\mathbb{R}$ as K and L go to infinity, and it is beyond the scope of this chapter to establish selection rules for these gridpoints.

Table (1.1) reports the results in the exogenous linear regression example. The number

[13] When K is odd, one of the equally spaced points is exactly equal to 0.5, and and the inverse standard CDF evaluated at 0.5 is equal to 0. Using 0 as a gridpoint is ruled out, since $e^{i \cdot 0 \cdot X} = 1$ is constant and cannot have a non-zero covariance with other random variables. We shift the equally spaced gridpoints by 0.01 to fix this issue.

of replications is set to 1000. When ϵ is normally distributed, the ordinary least squares estimator is already efficient since we assumed normality of the error term. We nevertheless show that our estimator does not perform much worse for most choices of K and L and in fact has a comparable MSE for the majority of choices of K and L. For larger K and L, we see the performance somewhat deteriorate, as expected from the finite properties of GMM with a large number of moments relative to N. For the mixture of normal distributions, we expect our estimator to yield asymptotic improvements over OLS, which is no longer efficient. For K and L between 2 and 5, the estimator based on independence exhibits a smaller bias, variance and other dispersion measures than OLS. Performance for $K = L = 1$ is very similar to OLS, and for $K = 6$ performance decreases. Performance of the independence-based estimator relative to OLS is slightly better for the larger sample size of $N = 500$.

When the distribution of ϵ has fatter tails, as is the case when it has a Student and Laplace distribution, the independence-based estimator has much smaller MSE when $K = L$ takes on values between 2 and 4. Again, when $K = L = 1$, the estimator's properties are very close to those of OLS and when $K = L = 6$, the MSE is larger than for other estimators considered here. We nevertheless see that for a judicious choice for K and L, we can get sizeable efficiency improvements from using the independence based-estimator when the distribution of ϵ is non-normal. When ϵ is normally distributed or when the choice of $K = L$ is sub-optimal, performance is not dramatically affected.

We now consider some linear instrumental variables model of the form $Y = X\beta_0 + \epsilon$ and $\epsilon \perp Z$. We compare the performance of the standard IV estimator $\beta_{IV} = (Z'X)^{-1}Z'Y$ to the estimator based on

$$E[e^{i\vec{t}_L(Y_i - X_i\theta)} \otimes e^{i\vec{s}_K Z_i}] - E[e^{i\vec{t}_L(Y_i - X_i\theta)}] \otimes E[e^{i\vec{s}_K Z_i}] = 0$$

for different choices of $K = L$, $\vec{s}_K$ and $\vec{t}_L$. We consider three designs with different levels of non-linearity and non-normality. All three impose $Y = X\beta_0 + \epsilon$ and $\beta_0 = 1$.

1. Normality: $\begin{pmatrix} X \\ \epsilon \\ Z \end{pmatrix} \sim N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix}$
2. Heavy tails: Z, ϵ and η are independently $t(5)$ and $X = Z(1 + \epsilon) + \sin(\epsilon)\cos(Z) + \eta$
3. Skewness: Z, ϵ and η are independent 50/50 mixtures of a normal distribution $N(-1, 1)$ and a normal distribution $N(1, 2)$. Furthermore, $X = Z + Z\epsilon + \eta$

Table (1.2) details the simulation results for these three designs with $N = 200$ and $N = 500$. The number of replications is set to 1000.

The IV estimator is efficient in the linear and normal design 1, and does perform better than the alternative, and more so for $N = 200$. The sample variance, IQR and MSE of the independence-based estimator is very close to that of the efficient IV, therefore performance

Table 1.1: Exogenous Linear Regression Monte Carlo

	N=200, ϵ with $N(0,1)$ distribution					N=500, ϵ with $N(0,1)$ distribution				
	Bias	S.Dev.	Median	IQR	R. MSE	Bias	S.Dev.	Median	IQR	R. MSE
OLS	-0.0030	0.074	0.0046	0.10	1.00	-0.0002	0.045	-0.0010	0.062	1.00
K=L=1	-0.0027	0.073	0.0043	0.10	0.98	-0.0002	0.045	-0.0010	0.062	1.00
K=L=2	-0.0023	0.075	0.0032	0.10	1.04	-0.0001	0.046	-0.0005	0.062	1.05
K=L=3	-0.0023	0.081	0.0025	0.11	1.18	-0.0006	0.047	-0.0010	0.063	1.10
K=L=4	-0.0023	0.085	0.0014	0.12	1.33	-0.0011	0.050	-0.0009	0.065	1.25
K=L=5	-0.0019	0.090	-0.0011	0.12	1.47	-0.0011	0.053	-0.0021	0.069	1.40
K=L=6	-0.0002	0.091	-0.0069	0.12	1.51	-0.0001	0.056	-0.0016	0.072	1.55
	N=200, ϵ with mixture distribution					N=500, ϵ with mixture distribution				
OLS	-0.0024	0.13	-0.0037	0.18	1.00	0.0008	0.085	0.0012	0.12	1.00
K=L=1	-0.0023	0.13	-0.0026	-0.79	1.01	0.0007	0.085	0.0007	0.12	1.00
K=L=2	0.0002	0.12	-0.0043	0.16	0.91	0.0025	0.079	-0.0011	0.11	0.88
K=L=3	0.0008	0.12	0.0018	0.16	0.84	0.0005	0.073	0.0003	0.10	0.75
K=L=4	0.0006	0.13	-0.0006	0.18	1.01	0.0002	0.076	-0.0005	0.10	0.81
K=L=5	0.0013	0.14	0.0007	0.18	1.12	0.0013	0.084	0.0011	0.11	0.97
K=L=6	0.0011	0.14	0.0003	0.20	1.21	0.0018	0.090	0.0009	0.12	1.13
	N=200, ϵ with Student distribution					N=500, ϵ with Student distribution				
OLS	-0.0028	0.12	-0.0003	0.16	1.00	0.0037	0.077	0.0037	0.10	1.00
K=L=1	-0.0027	0.12	-0.0001	0.16	0.99	0.0034	0.074	0.0033	0.10	0.92
K=L=2	-0.0005	0.10	0.0019	0.13	0.64	0.0034	0.058	0.0030	0.08	0.57
K=L=3	-0.0023	0.10	-0.0017	0.13	0.64	0.0031	0.057	0.0045	0.08	0.55
K=L=4	-0.0021	0.10	0.0004	0.14	0.75	0.0036	0.057	0.0028	0.08	0.55
K=L=5	-0.0041	0.11	-0.0028	0.15	0.86	0.0040	0.062	0.0016	0.08	0.63
K=L=6	-0.0029	0.12	-0.0033	0.15	1.04	0.0044	0.071	0.0009	0.09	0.83
	N=200, ϵ with Laplace distribution					N=500, ϵ with Laplace distribution				
OLS	0.0012	0.10	0.0002	0.14	1.00	0.0003	0.062	0.0014	0.083	1.000
K=L=1	0.0001	0.10	0.0010	0.14	1.01	-0.0002	0.062	0.0020	0.083	1.00
K=L=2	0.0016	0.09	0.0019	0.13	0.81	-0.0003	0.056	0.0021	0.077	0.82
K=L=3	0.0031	0.09	0.0099	0.12	0.84	-0.0005	0.055	0.0016	0.076	0.79
K=L=4	0.0015	0.10	0.0010	0.12	1.03	0.0005	0.058	0.0022	0.074	0.89
K=L=5	0.0022	0.11	0.0014	0.14	1.21	0.0013	0.069	0.0025	0.079	1.24
K=L=6	0.0032	0.12	0.0031	0.14	1.31	0.0011	0.074	0.0015	0.084	1.45

Notes: $Y = X\beta_0 + \epsilon$ with $\epsilon \perp X$, $X \sim N(0,1)$ and ϵ following the specified distribution. The mixture distribution is a 50/50 mixture combining $N(-1,1)$ and $N(1,4)$, two independent normal distributions. The Student distribution is has mean zero and 3 degrees of freedom. The Laplace has mean zero and scale parameter equal to 1. The number of replications is set to 1000. The first column contains the mean bias, the second contains the square root of the sampling variance of $\hat{\beta}$, the third contains the median of $\hat{\beta} - \beta_0$, the fourth contains the 75th quantile of $\hat{\beta}$ minus its 25th quantile, and the last column contains the relative MSE of the estimator vs. the OLS estimator's MSE.

Table 1.2: IV Linear Regression Monte Carlo

	N=200, Design 1					N=500, Design 1				
	Bias	S.Dev.	Median	IQR	R.MSE	Bias	S.Dev.	Median	IQR	R.MSE
IV	-0.006	0.10	-0.002	0.14	1.00	-0.001	0.064	0.000	0.080	1.00
K=L=1	-0.006	0.10	-0.003	0.14	1.00	-0.002	0.064	0.001	0.080	1.00
K=L=2	0.010	0.11	0.015	0.15	1.04	0.005	0.065	0.008	0.081	1.06
K=L=3	0.036	0.10	0.040	0.14	1.12	0.019	0.066	0.022	0.083	1.15
K=L=4	0.063	0.11	0.066	0.14	1.39	0.035	0.067	0.036	0.084	1.42
K=L=5	0.082	0.11	0.088	0.14	1.65	0.051	0.071	0.055	0.087	1.90
K=L=6	0.096	0.11	0.101	0.14	1.92	0.068	0.071	0.070	0.090	2.36
	N=200, Design 2					N=500, Design 2				
	Bias	S.Dev.	Median	IQR	R.MSE	Bias	S.Dev.	Median	IQR	R.MSE
IV	0.002	0.079	0.002	0.09	1.00	-0.001	0.047	-0.002	0.061	1.00
K=L=1	0.002	0.075	0.002	0.09	0.90	-0.001	0.047	-0.001	0.060	0.99
K=L=2	0.002	0.044	0.002	0.06	0.30	0.001	0.028	0.001	0.038	0.36
K=L=3	0.001	0.045	0.003	0.06	0.32	0.001	0.029	0.001	0.037	0.38
K=L=4	-0.001	0.052	-0.001	0.06	0.42	0.001	0.031	0.000	0.038	0.42
K=L=5	-0.002	0.055	-0.003	0.07	0.48	0.001	0.033	0.000	0.041	0.48
K=L=6	-0.003	0.061	-0.001	0.07	0.59	0.000	0.036	0.002	0.045	0.60
	N=200, Design 3					N=500, Design 3				
	Bias	S.Dev.	Median	IQR	R.MSE	Bias	S.Dev.	Median	IQR	R.MSE
IV	-0.001	0.074	0.002	0.096	1.00	-0.003	0.048	-0.002	0.066	1.00
K=L=1	-0.002	0.074	0.002	0.095	0.99	-0.003	0.048	-0.003	0.067	1.00
K=L=2	0.000	0.031	0.001	0.040	0.17	0.000	0.021	0.001	0.028	0.19
K=L=3	0.002	0.033	0.004	0.044	0.20	0.002	0.022	0.003	0.030	0.21
K=L=4	0.003	0.034	0.005	0.044	0.21	0.003	0.022	0.004	0.029	0.21
K=L=5	0.004	0.037	0.007	0.046	0.25	0.003	0.023	0.004	0.032	0.23
K=L=6	0.004	0.040	0.007	0.048	0.29	0.003	0.024	0.003	0.033	0.25

Notes: $Y = X\beta_0 + \epsilon$ with $\epsilon \perp Z$, $X \sim N(0,1)$ and ϵ, X follow distributions specified in the designs above. The number of replications is set to 1000. The first column contains the mean bias, the second contains the square root of the sampling variance of $\hat{\beta}$, the third contains the median of $\hat{\beta} - \beta_0$, the fourth contains the 75th quantile of $\hat{\beta}$ minus its 25th quantile, and the last column contains the relative MSE of the estimator vs. the IV estimator's MSE.

is not too adversely impacted by the choice of an equally efficient but more computationally challenging estimator. Measures of central tendency such as the bias and median deteriorate when using the independence-based estimator, and increasingly so as K and L are larger. This is a consequence of a bias term which is increasing in the number of moments used, but asymptotically goes to 0 if rates are picked appropriately.

In design 2, we show large improvements in the MSE, sample variance and IQR from using the independence-based estimate. There are also small bias and median improvements as well. The slow-decaying tails of the distributions of X and ϵ and the non-linearity in Z of $E[X|Z,\epsilon]$ both contribute to large efficiency improvements when using an estimator which is efficient for $\epsilon \perp Z$, rather than the IV estimator which is efficient under the weaker $\text{Cov}(Z,\epsilon) = 0$ restriction. The best performance coincides with a choice for K and L of 2 or 3, while again choosing $K = L = 1$ is not very different from choosing IV in terms of performance.

Finally, design 3 exhibits properties similar to design 2, except that the non-normality comes from the skewness of the distribution rather than the rate of decay for its tails. We again see 80% reductions in the MSE when K, L are between 2 and 5, and smaller biases for the majority of choices for K and L.

We also showcase the properties of the independence based estimator when the standard estimator is inconsistent. We consider two designs, the first being a linear regression $Y = X\beta_0 + \epsilon$, $\epsilon \perp X$ and ϵ is distributed along a standard Cauchy. The Cauchy's moments do not exist, therefore OLS will be inconsistent, but the estimator based on independence will be consistent and asymptotically normal. The second design is $Y = X\beta_0 + \epsilon$, $\epsilon \perp Z$ and ϵ has a standard Cauchy distribution. We compare our estimator to the standard linear IV, and show that mean-square error improvements are substantial.

1.6 Conclusion and Directions for Future Research

In this chapter we computed the efficiency bound for finite-dimensional parameters under various types of independence restrictions. We also proposed a GMM-type estimator which attains the efficiency bound, and performed a Monte Carlo study to study its finite-sample performance. There are several extensions of this chapter which could be interesting. The first would be to formalize the estimation results related to the estimation of models with unconditional independence restriction with nuisance functions, and to develop rates at which the sieve space for $F_0(\cdot)$ and the number of gridpoints in $\vec{s}_K$ and $\vec{t}_L$ must grow relative to the sample size N. Also, it would be useful to develop a rule for choosing K and L based on MSE calculations using higher-order expansions of the objective function, as in Donald et al. (2008). Such a study would allow one to better determine rates for K and L, and also to examine under what conditions the finite sample performance of the estimator will be satisfactory.

Table 1.3: Cauchy ϵ Monte Carlo

	Linear Regression with Cauchy ϵ				
	Bias	S. Dev.	Median	IQR	R.MSE
OLS	-0.441	8.55	0.054	1.56	1.000
K=L=1	0.009	0.29	0.002	0.28	0.001
K=L=2	-0.001	0.06	0.000	0.08	0.000
K=L=3	-0.003	0.06	-0.003	0.08	0.000
K=L=4	-0.001	0.05	-0.003	0.07	0.000
K=L=5	-0.001	0.05	-0.002	0.07	0.000
K=L=6	-0.002	0.05	-0.002	0.07	0.000
	Linear IV with Cauchy ϵ				
IV	-0.132	7.16	-0.011	1.56	1.000
K=L=1	-0.001	0.00	0.003	0.27	0.000
K=L=2	0.002	0.01	0.004	0.08	0.000
K=L=3	0.000	0.02	0.003	0.07	0.000
K=L=4	0.003	0.02	0.007	0.07	0.000
K=L=5	0.005	0.02	0.007	0.06	0.000
K=L=6	0.009	0.02	0.012	0.07	0.000

Notes: In the first design, $Y = X\beta_0 + \epsilon$ with $\epsilon \perp X$, $X \sim N(0,1)$ and ϵ is standard Cauchy. In the second design, $Y = X\beta_0 + \epsilon$ with $\epsilon \perp Z$, $Z \sim N(0,1)$, $X = Z + \arctan(\epsilon) + \eta$, $\eta \sim N(0,1)$ and ϵ is standard Cauchy. The number of replications is set to 1000. The first column contains the mean bias, the second contains the square root of the sampling variance of $\hat{\beta}$, the third contains the median of $\hat{\beta} - \beta_0$, the fourth contains the 75th quantile of $\hat{\beta}$ minus its 25th quantile, and the last column contains the relative MSE of the estimator vs. the simple estimator's MSE.

1.7 Proofs of Theorems and additional Lemmas

1.7.1 Efficiency Bound Calculations

Lemma 1.7.1 *The nuisance tangent space in model* (1.2) *is given by:*

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s}$$

where

$$\begin{aligned}
\Lambda_{1s} &= [a_1(W, \epsilon, X) : E[a_1(W, \epsilon, X)|\epsilon, X] = 0] \\
\Lambda_{2s} &= [a_2(\epsilon) : E[a_2(\epsilon)] = 0] \\
\Lambda_{3s} &= [a_3(X) : E[a_3(X)] = 0]
\end{aligned}$$

and these three subspaces are mutually orthogonal. Therefore projection of a zero-mean function h on Λ is:

$$\begin{aligned}
\Pi(h|\Lambda) &= \Pi(h|\Lambda_{1s}) + \Pi(h|\Lambda_{2s}) + \Pi(h|\Lambda_{3s}) \\
&= h - E[h|\epsilon, X] + E[h|\epsilon] + E[h|X]
\end{aligned}$$

Proof. The density of the observed variables, $f_{W,Y,X}(w, y, x|\theta)$ can be inverted to yield the density of (W, ϵ, X) as such:

$$\begin{aligned}
f_{W,Y,X}(y, z, x|\theta) &= |\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W,\epsilon,X}(w, \rho(y, w, \theta), x) \\
&= |\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W|\epsilon,X}(w|\rho(y, w, \theta), x)f_\epsilon(\rho(y, w, \theta))f_X(x)
\end{aligned}$$

Consider this parametric submodel with three additional parameters:

$$|\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W|\epsilon,X}(w|\rho(y, w, \theta), x, \gamma_1)f_\epsilon(\rho(y, w, \theta)|\gamma_2)f_X(x|\gamma_3)$$

and let γ_{10}, γ_{20} and γ_{30} denote the true values of the submodel parameters. The parametric submodel nuisance tangent spaces are, respectively:

$$\begin{aligned}
\Gamma_{\gamma_1} &= \{BS_{\gamma_1}(W, \epsilon, X) \text{ for all } B\} \\
S_{\gamma_1}(W, \epsilon, X) &= \frac{\partial}{\partial \gamma_1} \log f_{W|\epsilon,X}(W|\epsilon, X, \gamma_{10}) \\
\Gamma_{\gamma_2} &= \{BS_{\gamma_2}(\epsilon) \text{ for all } B\} \\
S_{\gamma_2}(\epsilon) &= \frac{\partial}{\partial \gamma_2} \log f_{\epsilon}(\epsilon|\gamma_{20}) \\
\Gamma_{\gamma_3} &= \{BS_{\gamma_3}(X) \text{ for all } B\} \\
S_{\gamma_3}(X) &= \frac{\partial}{\partial \gamma_3} \log f_{X}(X|\gamma_{30})
\end{aligned}$$

Using results from Tsiatis (2006) Chapter 4, we get that the mean-square closures of the nuisance tangent spaces are, respectively:

$$\begin{aligned}
\Lambda_{1s} &= [a_1(W, \epsilon, X) : E[a_1(W, \epsilon, X)|\epsilon, X] = 0] \\
\Lambda_{2s} &= [a_2(\epsilon) : E[a_2(\epsilon)] = 0] \\
\Lambda_{3s} &= [a_3(X) : E[a_3(X)] = 0]
\end{aligned}$$

We have Λ_{2s} orthogonal to Λ_{3s} by the independence of ϵ and X, and both Λ_{2s} and Λ_{3s} are orthogonal to Λ_{1s} by the fact that they are functions of the conditioning variables involved the definition of Λ_{1s}. The orthogonality of the subspaces allows us to write down the projection on the direct sum of the subspaces as the sum of the projections on the three subspaces. With Λ_{1s}, $\Pi(h|\Lambda_{1s}) = h - E[h|\epsilon, X]$ since for any $a_1(W, \epsilon, X) \in \Lambda_{1s}$, we have that:

$$\begin{aligned}
E[a_1(W, \epsilon, X)'(h - \Pi(h|\Lambda_{1s}))] &= E[a_1(W, \epsilon, X)'E[h|\epsilon, X]] \\
&= E[E[a_1(W, \epsilon, X)|\epsilon, X]'E[h|\epsilon, X]] \\
&= 0
\end{aligned}$$

For Λ_{2s}, $\Pi(h|\Lambda_{2s}) = E[h|\epsilon]$ since for any $a_2(\epsilon) \in \Lambda_{2s}$, we have that:

$$\begin{aligned} E[a_2(\epsilon)'(h - \Pi(h|\Lambda_{2s}))] &= E[a_2(\epsilon)'(h - E[h|\epsilon])] \\ &= E[a_2(\epsilon)'(E[h|\epsilon] - E[h|\epsilon])] \\ &= 0 \end{aligned}$$

The proof for Λ_{3s} is similar. ■

Proof of Lemma 1.2.2. The log-density of (Y, W, X) as a function of θ is equal to the following:

$$f_{W,Y,X}(W, Y, X|\theta) = |\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W,\epsilon,X}(W, \rho(Y, W, \theta), X)$$

and the score of the log-likelihood with respect to θ evaluated at θ_0 is:

$$\begin{aligned} S_\theta(W, \epsilon, X|\theta_0) &= \frac{\partial}{\partial \theta}\log|\frac{\partial}{\partial y'}\rho(Y, W, \theta)|_{\theta=\theta_0} + \frac{\partial}{\partial \theta}\rho(Y, W, \theta)|_{\theta=\theta_0}\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X}(W, \epsilon, X)}{f_{W,\epsilon,X}(W, \epsilon, X)} \\ &= J(Y, W, \theta_0) + \frac{\partial}{\partial \theta}\rho(Y, W, \theta)|_{\theta=\theta_0}\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X}(W, \epsilon, X)}{f_{W,\epsilon,X}(W, \epsilon, X)} \end{aligned}$$

where $J(Y, W, \theta_0) = \frac{\partial}{\partial \theta}\log|\frac{\partial}{\partial y'}\rho(Y, W, \theta)|_{\theta=\theta_0}$ is the Jacobian term. Projecting the score on the nuisance tangent space and computing the residual, we obtain the efficient score:

$$\begin{aligned} S^{\text{eff}}(X, \epsilon, \theta_0) &= E[S_\theta(W, \epsilon, X|\theta_0)|X, \epsilon] - E[S_\theta(W, \epsilon, X|\theta_0)|X] - E[S_\theta(W, \epsilon, X|\theta_0)|\epsilon] \\ &= E[J(Y, W, \theta_0) + \rho_\theta(Y, W, \theta_0)\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X}(W, \epsilon, X)}{f_{W,\epsilon,X}(W, \epsilon, X)}|X, \epsilon] - \\ &E[J(Y, W, \theta_0) + \rho_\theta(Y, W, \theta_0)\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X}(W, \epsilon, X)}{f_{W,\epsilon,X}(W, \epsilon, X)}|X] \\ &- E[J(Y, W, \theta_0) + \rho_\theta(Y, W, \theta_0)\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X}(W, \epsilon, X)}{f_{W,\epsilon,X}(W, \epsilon, X)}|\epsilon] \end{aligned}$$

$$E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}|X,\epsilon] = E[J(Y,W,\theta_0)|X,\epsilon]+$$
$$E[\rho_\theta(Y,W,\theta_0)|X,\epsilon]\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}+$$
$$\frac{\partial}{\partial\epsilon}E[\rho_\theta(Y,W,\theta_0)|X,\epsilon]$$
$$E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}|X] = 0$$
$$E[\frac{\partial}{\partial\theta}\rho(Y,W,\theta)|_{\theta=\theta_0}$$

$$\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}|\epsilon] = E[J(Y,W,\theta_0)|\epsilon]+E[\rho_\theta(Y,W,\theta_0)|\epsilon]\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$$
$$+\frac{\partial}{\partial\epsilon}E[\rho_\theta(Y,W,\theta_0)|\epsilon]$$

and therefore the efficient score is,

$$S^{\text{eff}}(X,\epsilon,\theta_0) = E[J(Y,W,\theta_0)|X,\epsilon]-E[J(Y,W,\theta_0)|\epsilon]+h(X,\epsilon)\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}+\frac{\partial}{\partial\epsilon}h(X,\epsilon),$$

with

$$h(X,\epsilon) = E[\rho_\theta(Y,W,\theta_0)|X,\epsilon]-E[\rho_\theta(Y,W,\theta_0)|\epsilon].$$

The efficient influence function and efficiency bound can both be computed directly using the efficient score. ■

Lemma 1.7.2 *The nuisance tangent space in model* (1.1) *is given by:*

$$\Lambda = \Lambda_{1s}\oplus\Lambda_{2s}\oplus\Lambda_{3s}\oplus\Lambda_{4s}$$

where

$$\Lambda_{1s} = [a_1(W, \epsilon, X, Z) : E[a_1(W, \epsilon, X, Z)|\epsilon, X, Z] = 0]$$
$$\Lambda_{2s} = [a_2(\epsilon, Z) : E[a_2(\epsilon, Z)|Z] = 0]$$
$$\Lambda_{3s} = [a_3(X, Z) : E[a_3(X, Z)|Z] = 0]$$
$$\Lambda_{4s} = [a_4(Z) : E[a_4(Z)] = 0]$$

and these four subspaces are mutually orthogonal. Therefore the projection of a zero-mean function h on Λ is:

$$\begin{aligned}\Pi(h|\Lambda) &= \Pi(h|\Lambda_{1s}) + \Pi(h|\Lambda_{2s}) + \Pi(h|\Lambda_{3s}) + \Pi(h|\Lambda_{4s}) \\ &= h - E[h|\epsilon, X, Z] + E[h|\epsilon, Z] + E[h|X, Z] - E[h|Z]\end{aligned}$$

Proof. The density of the observed variables, $f_{W,Y,X,Z}(w, y, x, z|\theta)$ can be inverted to yield the density of (W, ϵ, X, Z) as such:

$$\begin{aligned}f_{W,Y,X,Z}(w, y, z, x|\theta) &= |\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W,\epsilon,X,Z}(w, \rho(y, w, \theta), x, z) \\ &= |\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W|\epsilon,X,Z}(w|\rho(y, w, \theta), x, z)f_{\epsilon|z}(\rho(y, w, \theta)|z)f_{X|Z}(x|z)f_Z(z)\end{aligned}$$

Consider this parametric submodel with four additional parameters:

$$|\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W|\epsilon,X,Z}(z|\rho(y, z, \theta), x, w, \gamma_1)f_{\epsilon|Z}(\rho(y, z, \theta)|w, \gamma_2)f_{X|Z}(x|w, \gamma_3)f_Z(w|\gamma_4)$$

and let $\gamma_{10}, \gamma_{20}, \gamma_{30}$ and γ_{40} denote the true values of the submodel parameters. The parametric submodel nuisance tangent spaces are, respectively:

$$
\begin{aligned}
\Gamma_{\gamma_1} &= \{BS_{\gamma_1}(W, \epsilon, X, Z) \text{ for all } B\} \\
S_{\gamma_1}(W, \epsilon, X, Z) &= \frac{\partial}{\partial \gamma_1} \log f_{W|\epsilon,X,Z}(W|\epsilon, X, Z, \gamma_{10}) \\
\Gamma_{\gamma_2} &= \{BS_{\gamma_2}(\epsilon, Z) \text{ for all } B\} \\
S_{\gamma_2}(\epsilon, Z) &= \frac{\partial}{\partial \gamma_2} \log f_{\epsilon|Z}(\epsilon|Z, \gamma_{20}) \\
\Gamma_{\gamma_3} &= \{BS_{\gamma_3}(X, Z) \text{ for all } B\} \\
S_{\gamma_3}(X, Z) &= \frac{\partial}{\partial \gamma_3} \log f_{X|Z}(X|Z, \gamma_{30}) \\
\Gamma_{\gamma_4} &= \{BS_{\gamma_4}(Z) \text{ for all } B\} \\
S_{\gamma_4}(Z) &= \frac{\partial}{\partial \gamma_4} \log f_{Z}(Z|\gamma_{40})
\end{aligned}
$$

Using results from Tsiatis (2006) Chapter 4, we get that the mean-square closures of the nuisance tangent spaces are, respectively:

$$
\begin{aligned}
\Lambda_{1s} &= [a_1(W, \epsilon, X, Z) : E[a_1(W, \epsilon, X, Z)|\epsilon, X, Z] = 0] \\
\Lambda_{2s} &= [a_2(\epsilon, Z) : E[a_2(\epsilon, Z)|Z] = 0] \\
\Lambda_{3s} &= [a_3(X, Z) : E[a_3(X, Z)|Z] = 0] \\
\Lambda_{4s} &= [a_4(Z) : E[a_4(Z)] = 0]
\end{aligned}
$$

One can check directly that all these subspaces are mutually orthogonal. The orthogonality of the subspaces allows us to write down the projection on the direct sum of the subspaces as the sum of the projections on the four subspaces.

Zith Λ_{1s}, $\Pi(h|\Lambda_{1s}) = h - E[h|\epsilon, X, Z]$ since for any $a_1(W, \epsilon, X, Z) \in \Lambda_{1s}$, we have that:

$$
\begin{aligned}
E[a_1(W, \epsilon, X, Z)'(h - \Pi(h|\Lambda_{1s}))] &= E[a_1(W, \epsilon, X, Z)'E[h|\epsilon, X, Z]] \\
&= E[E[a_1(W, \epsilon, X, Z)|\epsilon, X, Z]'E[h|\epsilon, X, Z]] \\
&= 0
\end{aligned}
$$

For Λ_{2s}, $\Pi(h|\Lambda_{2s}) = E[h|\epsilon, Z] - E[h|Z]$ since for any $a_2(\epsilon, Z) \in \Lambda_{2s}$, we have that:

$$\begin{aligned}
E[a_2(\epsilon, Z)'(h - \Pi(h|\Lambda_{2s}))] &= E[a_2(\epsilon, Z)'(h - E[h|\epsilon, Z] + E[h|Z])] \\
&= E[a_2(\epsilon, Z)'(E[h|\epsilon, Z] - E[h|\epsilon, Z] + E[h|Z])] \\
&= E[a_2(\epsilon, Z)'E[h|Z]] \\
&= E[E[a_2(\epsilon, Z)|Z]'E[h|Z]] \\
&= 0
\end{aligned}$$

The proof for Λ_{3s} is similar, and that for Λ_{4s} is identical to that in the theorem concerning unconditional independence. ■

Proof of Lemma 1.2.4. The log-density of (Y, W, X, Z) as a function of θ is equal to the following:

$$f_{W,Y,X,Z}(W, Y, X, Z|\theta) = |\frac{\partial}{\partial y'}\rho(Y, W, \theta)|f_{W,\epsilon,X,Z}(W, \rho(Y, W, \theta), X, Z)$$

and the score of the log-likelihood with respect to θ evaluated at θ_0 is:

$$\begin{aligned}
S_\theta(W, \epsilon, X, Z|\theta_0) &= \frac{\partial}{\partial \theta}\log|\frac{\partial}{\partial y'}\rho(Y, W, \theta)|_{\theta=\theta_0} + \rho_\theta(Y, W, \theta_0)\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X,Z}(W, \epsilon, X, Z)}{f_{W,\epsilon,X,Z}(W, \epsilon, X, Z)} \\
&= J(Y, W, \theta_0) + \rho_\theta(Y, W, \theta_0)\frac{\frac{\partial}{\partial \epsilon}f_{W,\epsilon,X,Z}(W, \epsilon, X, Z)}{f_{W,\epsilon,X,Z}(W, \epsilon, X, Z)}
\end{aligned}$$

Zhere $J(Y, W, \theta_0) = \frac{\partial}{\partial \theta}\log|\frac{\partial}{\partial y'}\rho(Y, W, \theta)|_{\theta=\theta_0}$ is the Jacobian term. Projecting the score on the nuisance tangent space and computing the residual, we obtain the efficient score:

$$
\begin{aligned}
S^{\text{eff}}(W,X,\epsilon,Z) = {} & E[J(Y,W,\theta_0)+S_\theta(W,\epsilon,X,Z|\theta_0)|X,\epsilon,Z]- \\
& E[J(Y,W,\theta_0)+S_\theta(W,\epsilon,X,Z|\theta_0)|X,Z]- \\
& E[J(Y,W,\theta_0)+S_\theta(W,\epsilon,X,Z|\theta_0)|\epsilon,Z]+ \\
& E[J(Y,W,\theta_0)+S_\theta(W,\epsilon,X,Z|\theta_0)|Z] \\
= {} & E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|X,\epsilon,Z] \\
& - E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|X,Z] \\
& - E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|\epsilon,Z] \\
& + E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{Z,\epsilon,X,W}(Z,\epsilon,X,W)}{f_{Z,\epsilon,X,W}(Z,\epsilon,X,W)}|Z]
\end{aligned}
$$

$$
\begin{aligned}
E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|X,\epsilon,Z] = {} & E[J(Y,W,\theta_0)|X,\epsilon,Z] \\
& + \frac{\partial}{\partial\epsilon}E[\rho_\theta(Y,W,\theta_0)|X,\epsilon,Z] \\
& + \frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}(E[\rho_\theta(Y,W,\theta_0)|X,\epsilon,Z]
\end{aligned}
$$

$$
E[J(Y,W,\theta_0)+\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|X,Z] = 0
$$

$$
\begin{aligned}
E[\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|\epsilon,Z] = {} & E[J(Y,W,\theta_0)|\epsilon,Z] \\
& + \frac{\partial}{\partial\epsilon}E[\rho_\theta(Y,W,\theta_0)|\epsilon,Z] \\
& + \frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}(E[\rho_\theta(Y,W,\theta_0)|\epsilon,Z]
\end{aligned}
$$

$$
\begin{aligned}
E[\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|Z] &= E[E[\rho_\theta(Y,W,\theta_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}{f_{W,\epsilon,X,Z}(W,\epsilon,X,Z)}|X,Z]|Z] \\
&= E[0|Z] = 0
\end{aligned}
$$

and therefore the efficient score is,

$$S^{\text{eff}}(X,\epsilon,Z,\theta_0) = E[J(Y,W,\theta_0)|X,\epsilon,Z] - E[J(Y,W,\theta_0)|\epsilon,Z] + h(X,\epsilon,Z)\frac{f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)} + \frac{\partial}{\partial\epsilon}h(X,\epsilon,Z),$$

with

$$h(X,\epsilon,Z) = E[\rho_\theta(Y,W,\theta_0)|X,\epsilon,Z] - E[\rho_\theta(Y,W,\theta_0)|\epsilon,Z].$$

The efficient influence function and efficiency bound can both be computed directly using the efficient score.

■

Proof of Lemma 1.2.6. To prove the identification of θ_0, consider the conditional expectation of Y given $X = x$ and $Z = z$:

$$E[Y|X = x, Z = z] = x\theta_0 + E[G_0(Z,U)|X = x, Z = z].$$

If we assume that $U \perp X|Z$, we will have that

$$E[G_0(Z,U)|X = x, Z = z] = E[G_0(z,U)|Z = z],$$

and so $\theta_0 = \dfrac{\partial}{\partial x}E[Y|X = x, Z = z]$. If we assume $U \perp (X,Z)$, we have that $E[G_0(z,U)]$ does not depend on the value x, and therefore θ_0 is identified here as well.

To make use of Lemma (1.2.4), we will show that $U \perp X|Z$ if and only if $G(Z,U) = Y - X\theta_0 \perp X|Z$. Let $U \perp X|Z$. Then,

$$\begin{aligned} P(G_0(Z,U) < a|X = x, Z = z) &= P(G_0(z,U) < a|X = x, z = z) \\ &= P(G_0(z,U) < a|Z = z) \\ &= P(G_0(Z,U) < a|Z = z) \end{aligned}$$

so that $G_0(Z,U) \perp X|Z$ if $U \perp X|Z$. Now, assume that $G_0(Z,U) \perp X|Z$, which implies that:

$$\begin{aligned} P(U < a|X = x, Z = z) &= P(G_0(Z, U) < G_0(z, a)|X = x, Z = z) \\ &= P(G_0(z, Y) < G_0(z, a)|Z = z) \\ &= P(Y < a|Z = z) \\ \Rightarrow & U \perp X|Z. \end{aligned}$$

We can now apply Lemma (1.2.4) and see that the efficiency bound is the one presented in Lemma (1.2.6). Now, to show that the efficiency bound does not differ when we assume $U \perp (X, Z)$, we will use the following fact:

$$(U \perp X|Z \text{ and } U \perp Z) \Leftrightarrow U \perp (X, Z).$$

Define $\tilde{U} = \Phi^{-1}(F_{U|Z}(U|Z))$, where $F_{U|Z}(\cdot)$ is the conditional CDF of U given Z, and Φ is the CDF of a standard normal distribution. Since U is continuously distributed given Z, we have that $\tilde{U}$ is $N(0, 1)$ distributed, also independently from Z. Let $\tilde{G}(Z, a) = G(Z, (F_{U|Z}(\Phi(a)|Z)))$. $\tilde{G}(Z, \tilde{U}) = G(Z, U)$, and $\tilde{G}(\cdot)$ also satisfies Assumption (1.2.5) (a). $\tilde{U} \perp Z$ by construction and we also have that $\tilde{U} \perp X|Z$, which implies that $\tilde{U} \perp (X, Z)$. We have therefore shown that we can renormalize the function $G(\cdot)$ so that $U \perp Z$ without violating any of the assumptions necessary for this lemma, therefore, we can assume that $U \perp Z$ without loss of generality. ■

Lemma 1.7.3 *The nuisance tangent space in model* (1.3) *is given by:*

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}$$

where

$$\begin{aligned} \Lambda_{1s} &= [a_1(Y, X) : E[a_1(Y, X)|Y, V] = 0] \\ \Lambda_{2s} &= [a_2(X) : E[a_2(X)] = 0] \end{aligned}$$

and these subspaces are mutually orthogonal. Therefore the projection of a zero-mean function h on Λ is:

$$\begin{aligned}\Pi(h|\Lambda) &= \Pi(h|\Lambda_{1s}) + \Pi(h|\Lambda_{2s}) \\ &= E[h|Y,V] - E[h|V] + E[h|X] + E[h]\end{aligned}$$

Proof. The density of the observed variables (Y, X) has the following decomposition:

$$\begin{aligned}f_{Y,X}(Y,X) &= f_{Y|X}(Y|X)f_X(X) \\ &= f_{Y|X,V}(Y|X,V(X,\theta))f_X(X) \\ &= f_{Y|V}(Y|V(X,\theta))f_X(X)\end{aligned}$$

Consider this parametric submodel with parameters γ_1 and γ_2

$$f_{Y|V}(Y|V(X,\theta),\gamma_1)f_X(X,\gamma_2)$$

and let γ_{10} and γ_{20} denote the true value of the submodel parameters. The parametric submodel nuisance tangent spaces are, respectively:

$$\begin{aligned}\Gamma_{\gamma_1} &= \{BS_{\gamma_1}(Y,X) \text{ for all } B\} \\ S_{\gamma_1}(Y,X) &= \frac{\partial}{\partial\gamma_1}\log f_{Y|V}(Y|V(X,\theta_0),\gamma_{10}) \\ \Gamma_{\gamma_2} &= \{BS_{\gamma_2}(X) \text{ for all } B\} \\ S_{\gamma_2}(X) &= \frac{\partial}{\partial\gamma_2}\log f_X(X|\gamma_{20})\end{aligned}$$

Using results from Tsiatis (2006) Chapter 4, we get that the mean-square closures of the nuisance tangent spaces are, respectively:

$$\begin{aligned}\Lambda_{1s} &= [a_1(Y,X) : E[a_1(Y,X)|V] = 0 \text{ and } a_1(\cdot) \text{ depends on } X \text{ only through } V] \\ \Lambda_{2s} &= [a_2(X) : E[a_2(X)] = 0]\end{aligned}$$

One can check directly that all these subspaces are orthogonal. The orthogonality of the subspaces allows us to write down the projection on the direct sum of the subspaces as the

sum of the projections on both subspaces.

With Λ_{1s}, $\Pi(h|\Lambda_{1s}) = E[h|Y,V] - E[h|V]$ since for any $a_1(Y,V) \in \Lambda_{1s}$, we have that:

$$\begin{aligned}
E[a_1(Y,V)'(h - \Pi(h|\Lambda_{1s}))] &= E[a_1(Y,V)'(h - E[h|Y,V] + E[h|V])] \\
&= E[a_1(Y,V)'(E[h|Y,V] - E[h|Y,V] + E[h|V])] \\
&= E[a_1(Y,V)'E[h|V]] \\
&= E[E[a_1(Y,V)|V]'E[h|V]] \\
&= 0,
\end{aligned}$$

and from previous results $\Pi(h|\Lambda_{2s}) = E[h|X] - E[h]$. ■

Proof of Lemma 1.3. The log-density of (Y,X) as a function of θ is equal to the following:

$$\log f_{Y,X}(Y,X|\theta) = \log f_{Y|V}(Y|V(X,\theta)) + \log f_X(X)$$

and the score of the log-likelihood with respect to θ evaluated at θ_0 is:

$$S_\theta(Y,X|\theta_0) = V_\theta(X,\theta_0)\frac{\frac{\partial}{\partial V}f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)}$$

Projecting the score on the nuisance tangent space, we obtain the efficient score:

$$\begin{aligned}
S^{\text{eff}}(Y,X,\theta_0) &= S_\theta(Y,X|\theta_0) - E[S_\theta(Y,X|\theta_0)|Y,V] + \\
&\quad E[S_\theta(Y,X|\theta_0)|V] - E[S_\theta(Y,X|\theta_0)|X] \\
E[S_\theta(Y,X|\theta_0)|Y,V] &= E[V_\theta(X,\theta_0)|Y,V]\frac{\frac{\partial}{\partial V}f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)} \\
&= E[V_\theta(X,\theta_0)|V]\frac{\frac{\partial}{\partial V}f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)}
\end{aligned}$$

where the last equality comes from $X \perp Y|V$.

$$
\begin{aligned}
E[S_\theta(Y,X|\theta_0)|X] &= V_\theta(X,\theta_0)E\left[\frac{\frac{\partial}{\partial V}f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)}|X\right] \\
&= V_\theta(X,\theta_0)\int \frac{\frac{\partial}{\partial V}f_{Y|V}(y|V)}{f_{Y|V}(y|V)}f_{Y|X}(y|X)dy \\
&= V_\theta(X,\theta_0)\int \frac{\frac{\partial}{\partial V}f_{Y|V}(y|V)}{f_{Y|V}(y|V)}f_{Y|V}(y|V)dy \\
&= V_\theta(X,\theta_0)\int \frac{\partial}{\partial V}f_{Y|V}(y|V)dy \\
&= V_\theta(X,\theta_0)\frac{\partial}{\partial V}\int f_{Y|V}(y|V)dy \\
&= 0 \\
E[S_\theta(Y,X|\theta_0)|V] &= E[E[S_\theta(Y,X|\theta_0)|X,V]|V] \\
&= E[E[S_\theta(Y,X|\theta_0)|X]|V] \\
&= 0.
\end{aligned}
$$

Therefore, the efficient score in this model is:

$$
S^{\text{eff}}(Y,X,\theta_0) = (V_\theta(X,\theta_0) - E[V_\theta(X,\theta_0)|V])\frac{\frac{\partial}{\partial V}f_{Y|V}(Y|V)}{f_{Y|V}(Y|V)},
$$

and the efficient influence function and variance bound can be computed directly using the efficient score, completing the proof.

■

Lemma 1.7.4 *The nuisance tangent space in model* (1.4) *is given by:*

$$
\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s} \oplus \Lambda_{4s}
$$

where

$$\Lambda_{1s} = [a_1(W,\epsilon,X) : E[a_1(W,\epsilon,X)|\epsilon,X] = 0]$$
$$\Lambda_{2s} = [a_2(\epsilon) : E[a_2(\epsilon)] = 0]$$
$$\Lambda_{3s} = [a_3(X) : E[a_3(X)] = 0]$$
$$\Lambda_{4s} = \left[a_4(W,\epsilon,X) = \log|\frac{\partial}{\partial y'}\rho(Y,W,\alpha_0)|[w] + \frac{\frac{\partial}{\partial \epsilon} f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}\rho(Y,W,\alpha_0)[w], w \in \mathcal{F} \right]$$

and the first three subspaces are mutually orthogonal. Therefore, the projection of a zero-mean function h on Λ is:

$$\Pi(h|\Lambda) = h - \Delta[w^*] - E[h - \Delta[w^*]|X,\epsilon] + E[h - \Delta[w^*]|\epsilon] + E[h - \Delta[w^*]|X]$$

where

$$\Delta[w] = \log|\frac{\partial}{\partial y'}\rho(Y,W,\alpha_0)|[w] + \frac{\frac{\partial}{\partial \epsilon} f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}\rho(Y,W,\alpha_0)[w]$$
$$w^* = argmin_{w\in\mathcal{F}} E\left[(E[h - \Delta[w]|X,\epsilon] - E[h - \Delta[w]|\epsilon] - E[h - \Delta[w]|X])^2\right]$$

Proof. The density of the observed variables, $f_{W,Y,X}(w,y,x|\alpha)$ can be inverted to yield the density of (W,ϵ,X) as such:

$$\begin{aligned} f_{W,Y,X}(w,y,x|\alpha) &= |\frac{\partial}{\partial y'}\rho(Y,W,\alpha)| f_{W,\epsilon,X}(w,\rho(y,w,\alpha),x) \\ &= |\frac{\partial}{\partial y'}\rho(Y,W,\alpha)| f_{W|\epsilon,X}(w|\rho(y,w,\alpha),x) f_\epsilon(\rho(y,w,\alpha)) f_X(x) \end{aligned}$$

Consider this parametric submodel with four additional parameters:

$$|\frac{\partial}{\partial y'}\rho(Y,W,\theta,F(\cdot,\gamma_4))| f_{W|\epsilon,X}(w|\rho(y,w,\theta,F(\cdot,\gamma_4)),x,\gamma_1) f_\epsilon(\rho(y,w,\theta,F(\cdot,\gamma_4))|\gamma_2) f_X(x|\gamma_3)$$

and let γ_{10}, γ_{20}, γ_{30} and γ_{40} denote the true values of the submodel parameters. The parametric submodel nuisance tangent spaces are, respectively:

$$\begin{aligned}
\Gamma_{\gamma_1} &= \{BS_{\gamma_1}(W, \epsilon, X) \text{ for all } B\} \\
S_{\gamma_1}(W, \epsilon, X) &= \frac{\partial}{\partial \gamma_1} \log f_{W|\epsilon,X}(W|\epsilon, X, \gamma_{10}) \\
\Gamma_{\gamma_2} &= \{BS_{\gamma_2}(\epsilon) \text{ for all } B\} \\
S_{\gamma_2}(\epsilon) &= \frac{\partial}{\partial \gamma_2} \log f_{\epsilon}(\epsilon|\gamma_{20}) \\
\Gamma_{\gamma_3} &= \{BS_{\gamma_3}(X) \text{ for all } B\} \\
S_{\gamma_3}(X) &= \frac{\partial}{\partial \gamma_3} \log f_{X}(X|\gamma_{30}) \\
\Gamma_{\gamma_4} &= \{BS_{\gamma_4}(W, \epsilon, X) \text{ for all } B\} \\
S_{\gamma_4}(W, \epsilon, X) &= \Delta[w].
\end{aligned}$$

Using results from Tsiatis (2006) Chapter 4, we get that the mean-square closures of the nuisance tangent spaces are, respectively:

$$\begin{aligned}
\Lambda_{1s} &= [a_1(W, \epsilon, X) : E[a_1(W, \epsilon, X)|\epsilon, X] = 0] \\
\Lambda_{2s} &= [a_2(\epsilon) : E[a_2(\epsilon)] = 0] \\
\Lambda_{3s} &= [a_3(X) : E[a_3(X)] = 0] \\
\Lambda_{4s} &= [\Delta[w] : w \in \mathcal{F}]
\end{aligned}$$

Using previous results, we see that Λ_{1s}, Λ_{2s} and Λ_{3s} are mutually orthogonal. To show that

$$\Pi(h|\Lambda) = h - \Delta[w^*] - E[h - \Delta[w^*]|X, \epsilon] + E[h - \Delta[w^*]|\epsilon] + E[h - \Delta[w^*]|X],$$

we will use the fact that the projection minimizes the expectation of the squared distance between h and the projection of h onto Λ. We choose $(a_1(\cdot), a_2(\cdot), a_3(\cdot), \Delta[\cdot]) \in \times_{i=1}^{4} \Lambda_{is}$ such that

$$E\left[(h - \Delta[w] - a_1(W, \epsilon, X) - a_2(\epsilon) - a_3(X))^2\right]$$

is minimized. The minimization over $\times_{i=1}^{4}\Lambda_{is}$ can be done sequentially, such that $h-\Delta[w]$ is projected on $\times_{i=1}^{3}\Lambda_{is}$ for any $w\in\mathcal{F}$, and then this projection is further projected on Λ_{4s}. Using previous results, the projection of $h-\Delta[w]$ on $\times_{i=1}^{3}\Lambda_{is}$ is

$$h-\Delta[w]-E[h-\Delta[w]|X,\epsilon]+E[h-\Delta[w]|\epsilon]+E[h-\Delta[w]|X],$$

and therefore,

$$w^{*}=\text{argmin}_{w\in\mathcal{F}}E\left[(h-\Delta[w]-E[h-\Delta[w]|X,\epsilon]+E[h-\Delta[w]|\epsilon]+E[h-\Delta[w]|X])^{2}\right].$$

■

Proof of Lemma 1.2.10. The log-density of (Y,W,X) as a function of α is equal to the following:

$$f_{W,Y,X}(W,Y,X|\theta)=|\frac{\partial}{\partial y'}\rho(Y,W,\alpha)|f_{W,\epsilon,X}(W,\rho(Y,W,\alpha),X)$$

and the score of the log-likelihood with respect to θ evaluated at α_0 is:

$$\begin{aligned}S_{\theta}(W,\epsilon,X|\alpha_0)&=\frac{\partial}{\partial\theta}\log|\frac{\partial}{\partial y'}\rho(Y,W,\alpha)|_{\alpha=\alpha_0}+\rho_{\theta}(Y,W,\alpha_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}\\&=J(Y,W,\alpha_0)+\rho_{\theta}(Y,W,\alpha_0)\frac{\frac{\partial}{\partial\epsilon}f_{W,\epsilon,X}(W,\epsilon,X)}{f_{W,\epsilon,X}(W,\epsilon,X)}\end{aligned}$$

where $J(Y,W,\theta_0)=\frac{\partial}{\partial\theta}\log|\frac{\partial}{\partial y'}\rho(Y,W,\alpha)|_{\alpha=\alpha_0}$ is the Jacobian term. Projecting the score on the nuisance tangent space and computing the residual, we obtain the efficient score:

$$\begin{aligned}S^{\text{eff}}(X,\epsilon,\alpha_0)&=E[S_{\theta}(W,\epsilon,X|\alpha_0)-\Delta[w^{*}]|X,\epsilon]-E[S_{\theta}(W,\epsilon,X|\theta_0)-\Delta[w^{*}]|X]\\-E[S_{\theta}(W,\epsilon,X|\theta_0)-\Delta[w^{*}]|\epsilon]&-E[S_{\theta}(W,\epsilon,X|\theta_0)-\Delta[w^{*}]|X]=0\end{aligned}$$

and therefore the efficient score is,

$$
\begin{aligned}
S^{\text{eff}}(X,\epsilon,\alpha_0) &= E[S_\theta(W,\epsilon,X|\alpha_0) - \Delta[w^*]|X,\epsilon] - E[S_\theta(W,\epsilon,X|\theta_0) - \Delta[w^*]|\epsilon] \\
&= E[S_\theta(W,\epsilon,X|\alpha_0) + S_F[w^*] + J[w^*]|X,\epsilon] - \\
&\quad E[S_\theta(W,\epsilon,X|\alpha_0) + S_F[w^*] + J[w^*]|\epsilon].
\end{aligned}
$$

The efficient influence function and efficiency bound can both be computed directly using the efficient score. ■

Proof of lemma (1.3.3)**.** ($\Leftarrow$): If X is independent from Y, then $f(X)$ is independent from $g(Y)$ for any functions f, g. Since independence implies uncorrelatedness, we are done. ($\Rightarrow$): Let $Z = (X,Y)$, then $\mathcal{H} = \mathcal{F} \times \mathcal{G}$ is a convergence determining class for Z. Let $\tilde{Z} = (\tilde{X}, \tilde{Y})$ where $\tilde{X}$ (and $\tilde{Y}$) has the same distribution as X (and Y), but with $\tilde{X} \perp \tilde{Y}$. Then,

$$
\begin{aligned}
&\text{Cov}\,(f(X), g(Y)) = 0 \ \text{ for all } (f,g) \in \mathcal{F} \times \mathcal{G} \\
&\Rightarrow E[f(X)g(Y)] = E[f(X)]E[g(Y)] \ \text{ for all } (f,g) \in \mathcal{F} \times \mathcal{G} \\
&\Rightarrow E[h(Z)] = E[h(\tilde{Z})] \text{ for all } h \in \mathcal{H} \\
&\Rightarrow Z \stackrel{d}{=} \tilde{Z}
\end{aligned}
$$

Therefore, $Z = (X,Y)$ has independent marginal distributions. ■

1.7.2 Consistency under Unconditional Independence

Lemma 1.7.5 *Let assumption 1.3.1 hold. Without loss of generality, we can assume that* $E[q_i^K] = 0$, $E[q_i^K q_i^{K\prime}] = I_K$ *and* $\sup_{x \in \mathcal{X}} \|q^K(x)\| \leq C\sqrt{K}$. *We can also assume that,* $E[\rho^L(\theta)] = 0 \forall \theta$, $E[\rho_i^L(\theta_0)\rho_i^L(\theta_0)'] = I_L$ *and* $\sup_{(w,y) \in \mathcal{W} \times \mathcal{Y}} \|\rho^L(\theta_0)(y,w)\| \leq C\sqrt{L}$.

Proof. Note that the projection based estimators are invariant to a linear transformation of the r_i^K. Letting $B_1 = E[r_i^K r_i^{K\prime}]^{-1/2}$, $E[B_1 r_i^K (B_1 r_i^K)'] = I_K$, and $sup_{x \in X} \|B_1 r_i^K\| \leq C\sqrt{K}$ where C is the inverse of the smallest eigenvalue of $E[r_i^K r_i^{K\prime}]$, which is bounded away from 0 by assumption. Letting $B_2 = (I_K - E[r_i^K]E[r_i^K]')^{-1/2}$, $E[B_2 B_1 q_i^K (B_2 B_1 q_i^K)'] = I_K$. The proof for ρ^L is similar. ■

Let $R(\theta) = \|E[g(\cdot,\cdot,\theta)]\|_\Omega^2$, where $g(s,t,\theta) = \text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}\right)$ and $\|g\|_\Omega$ denotes the reproducing kernel Hilbert space norm of g. See Parzen (1959) for more details on Hilbert spaces and reproducing kernel spaces.

Lemma 1.7.6 (Identification) $R(\theta) = 0 \Rightarrow \theta = \theta_0$

Proof.

$$
\begin{aligned}
R(\theta) = 0 &\Rightarrow \Omega^{-1/2} E[g(\cdot,\cdot,\theta)] = 0 \\
&\Leftrightarrow E[g(\cdot,\cdot,\theta)] = 0 \text{ Since } \Omega^{-1/2} \text{ is an invertible operator} \\
&\Leftrightarrow \text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}\right) = 0 \text{ for all } s,t \in \mathbb{R} \\
&\Leftrightarrow \rho(Y,W,\theta) \perp X \text{ since complex exponentials are a convergence determining class} \\
&\Leftrightarrow \theta = \theta_0 \text{ from the identifying assumption.}
\end{aligned}
$$

■

Lemma 1.7.7 (Continuity) *$R(\theta)$ is continuous in θ for all $\theta \in \tilde{\Theta}$, a compact subset of the parameter space Θ.*

Proof.

Let $\tilde{\Theta} = \Theta \cap \{\theta : R(\theta) \leq C\}$, where C is a constant such that $0 < C < \infty$. This restriction is imposed to avoid having an infinite valued population objective function. Constraining the parameter space in such a way is not restrictive, because it eliminates parameters for which the population objective is extremely large. We rule out that $R(\theta) = \infty$ when $\theta \neq \theta_0$ later through an assumption about the conditional density of ϵ and the residual function $\rho(\cdot)$.

We have that

$$
\begin{aligned}
R(\theta) &= \|\text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}\right)\|_\Omega^2 \\
\text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}\right) &= E[e^{is\rho(Y,W,\theta)} e^{itX}] - E[e^{is\rho(Y,W,\theta)}] E[e^{itX}] \\
&= E[e^{is\rho(m(\epsilon,W,\theta_0),W,\theta)} e^{itX}] - E[e^{is\rho(m(\epsilon,W,\theta_0),W,\theta)}] E[e^{itX}]
\end{aligned}
$$

$m(\cdot,W,\theta)$ is the inverse function of $\rho(\cdot,W,\theta)$, which exists and is differentiable by assumption (1.3.4)(d).

$$
E[e^{is\rho(m(\epsilon,W,\theta_0),W,\theta)}|X,W] = \int e^{is\rho(m(\epsilon,W,\theta_0),W,\theta)} f(\epsilon|X,W) d\epsilon
$$

Let $u = \rho(m(\epsilon, W, \theta_0), W, \theta)$, which implies that $\epsilon = \rho(m(u, W, \theta), W, \theta_0)$. Also, we see that

$$\begin{aligned} du &= \frac{\partial}{\partial \epsilon}\rho(m(\epsilon, W, \theta_0), W, \theta)d\epsilon \\ &= g(\epsilon, W, \theta)d\epsilon \\ &= g(\rho(m(u, W, \theta), W, \theta_0), W, \theta)d\epsilon. \end{aligned}$$

Also note that $g(\epsilon, W, \theta_0) = 1$. Therefore,

$$\begin{aligned} E[e^{is\rho(m(\epsilon,W,\theta_0),W,\theta)}|X, W] &= \int \frac{e^{isu} f_{\epsilon|X,W}(\rho(m(W, u, \theta), W, \theta_0)|X, W)}{g(\rho(m(u, W, \theta), W, \theta_0), W, \theta)}du \\ &= \int e^{isu}\frac{f\epsilon|X, W(\rho(m(W, u, \theta), W, \theta_0)|X, W)}{f\epsilon|X, W(u|X, W)} \times \\ &\quad \frac{f\epsilon|X, W(u|X, W)}{g(\rho(m(u, W, \theta), W, \theta_0), W, \theta)}du \\ &\quad \text{since the conditional density of } \epsilon \text{ is non-zero} \\ &= E[e^{is\epsilon}\frac{f_{\epsilon|X,W}(\rho(m(W, \epsilon, \theta), W, \theta_0)|X, W)}{f_{\epsilon|X,W}(\epsilon|X, W)} \times \\ &\quad \frac{1}{g(\rho(m(\epsilon, W, \theta), W, \theta_0), W, \theta)}|X, W]. \end{aligned}$$

Let

$$\begin{aligned} U(X, \epsilon, \theta) &= E[\frac{f_{\epsilon|X,W}(\rho(m(W, \epsilon, \theta), W, \theta_0)|X, W)}{g(\rho(m(\epsilon, W, \theta), W, \theta_0), W, \theta)f_{\epsilon|X,W}(\epsilon|X, W)}|X, \epsilon] \\ \text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}\right) &= E[U(X, \epsilon, \theta)e^{is\epsilon}e^{itX}] - E[U(X, \epsilon, \theta)e^{is\epsilon}]E[e^{itX}] \end{aligned}$$

Using an argument similar to that in the proof of the efficiency bound equivalence, we see that:

$$R(\theta) = E[(U(X, \epsilon, \theta) - E[U(X, \epsilon, \theta)|\epsilon])^2].$$

By assumption (1.3.4), we have that $E[U(X, W, \epsilon)^2] < \infty$ on a neighborhood of θ_0. To show continuity, see that by assumptions (1.3.4)(c)-(e) $f_{\epsilon|X,W}(\rho(m(W, \epsilon, \theta), W, \theta_0)|X, W)$ is continuous in θ, and so is $\frac{1}{g(\rho(m(\epsilon, W, \theta), W, \theta_0), W, \theta)}$. Therefore, $U(X, \epsilon, \theta)$ is continuous in θ, and $R(\theta)$ is continuous in θ on when $R(\theta) < \infty$. Note that this implies that $\tilde{\Theta} = \Theta \cap \{\theta :$

$R(\theta) \leq C\}$ is compact, since $\{\theta : R(\theta) \leq C\}$ is closed by the continuity of $R(\cdot)$ and bounded by picking C small enough. ■

Let $\Omega^{KL}(\theta_0) = E[\rho_i^L(\theta_0)\rho_i^L(\theta_0)'] \otimes E[q_i^K q_i^{K\prime}]$ and $\widehat{\widehat{\Omega}}^{KL}(\tilde{\theta}) = \frac{1}{N}\sum_{i=1}^N \hat{\rho}_i^L(\tilde{\theta})\hat{\rho}_i^L(\tilde{\theta})' \otimes \hat{q}_i^K \hat{q}_i^{K\prime}$. Also, let $\|A\|_2 = \sqrt{\text{tr}(A^2)}$ denote the Hilbert-Schimdt norm and $\|A\|_\infty = \lambda_{\max}(A)$ be the spectral norm of A, where $\lambda_{\max}(A)$ is the largest eigenvalue of A, a symmetric matrix.

Lemma 1.7.8 *Let $\rho^L(\theta)$ and q^K be basis functions that satisfy assumption* (1.3.1). *Let $\tilde{\theta} - \theta_0 = O_p(\tau_N)$, then $\|\widehat{\widehat{\Omega}}^{KL}(\tilde{\theta}) - \Omega^{KL}(\theta_0)\|_2 \xrightarrow{\mathbb{P}} 0$ if $KL\tau_N \to 0$, $\sqrt{\frac{KL}{N}}\sqrt{K}\sqrt{L} \to 0$ and $\sqrt{\frac{KL}{N}}\frac{\sqrt{K}+\sqrt{L}}{N} \to 0$ as $N \to \infty$. Also, $\|\tilde{\widehat{\Omega}}^{KL}(\tilde{\theta})^{-1} - \Omega^{KL}(\theta_0)^{-1}\| \xrightarrow{\mathbb{P}} 0$ if if the same rates for K and L are chosen.*

Proof. Let

$$\begin{aligned}
\Omega_1 &= \Omega^{KL}(\theta_0) \\
&= E[\rho_i^L(\theta_0)\rho_i^L(\theta_0)'] \otimes E[q_i^K q_i^{K\prime}] \\
\Omega_2 &= \frac{1}{N}\sum_{i=1}^N \rho_i^L(\theta_0)\rho_i^L(\theta_0)' \otimes q_i^K q_i^{K\prime} \\
\Omega_3 &= \frac{1}{N}\sum_{i=1}^N \rho_i^L(\theta_0)\rho_i^L(\theta_0)' \otimes \hat{q}_i^K \hat{q}_i^{K\prime} \\
\Omega_4 &= \frac{1}{N}\sum_{i=1}^N \hat{\rho}_i^L(\theta_0)\hat{\rho}_i^L(\theta_0)' \otimes \hat{q}_i^K \hat{q}_i^{K\prime} \\
\Omega_5 &= \widehat{\widehat{\Omega}}^{KL}(\tilde{\theta}) = \frac{1}{N}\sum_{i=1}^N \hat{\rho}_i^L(\tilde{\theta})\hat{\rho}_i^L(\tilde{\theta})' \otimes \hat{q}_i^K \hat{q}_i^{K\prime}
\end{aligned}$$

By lemma (1.7.5), we assume without loss of generality

$$\begin{aligned}
E[\|\Omega_2 - \Omega_1\|_2^2] &= E[\|\frac{1}{N}\sum_{i=1}^N \rho_i^L(\theta_0)\rho_i^L(\theta_0)' \otimes q_i^K(q_i^K)' - E[\rho_i^L(\theta_0)\rho_i^L(\theta_0)'] \otimes E[q_i^K(q_i^K)']\|^2] \\
&= E[\|\frac{1}{N}\sum_{i=1}^N A_i \otimes B_i - E[A_i] \otimes E[B_i]\|^2]
\end{aligned}$$

where $A_i = \rho_i^L(\theta_0)\rho_i^L(\theta_0)'$, and WLOG, $E[A_i] = I_L$. Similarly, $B_i = q_i^K q_i^{K\prime}$ and $E[B_i] = I_K$. Let $C_i = A_i \otimes B_i - E[A_i] \otimes E[B_i]$. Using independence of A_i and B_i, we have that $E[C_i] = 0$. Therefore:

$$
\begin{aligned}
E[\|\frac{1}{N}\sum_{i=1}^{N} C_i\|_2^2] &= \frac{1}{N^2}\mathrm{tr}(\sum_{i=1}^{N} E[C_i^2] + \sum_{i<j} E[C_iC_j]) \\
&= \frac{1}{N}\mathrm{tr}E[C_i^2] \\
&= \frac{1}{N}\mathrm{tr}(E[A_i^2 \otimes B_i^2] - E[A_i]^2 \otimes E[B_i]^2) \\
&= \frac{1}{N}\mathrm{tr}(E[A_i^2] \otimes E[B_i^2] - I_{KL}) \\
&= \frac{1}{N}\mathrm{tr}(E[A_i^2])\mathrm{tr}(E[B_i^2]) - \frac{KL}{N} \\
&= \frac{1}{N}E[\|\rho_i^L(\theta_0)\|^4]E[\|q_i^K\|^4] - \frac{KL}{N} \\
&\leq \frac{1}{N}\sqrt{L}^2 L\sqrt{K}^2 K - \frac{KL}{N} \\
&\leq \frac{1}{N}K^2L^2
\end{aligned}
$$

which implies, by Markov's inequality, that $\|\Omega_2 - \Omega_1\|_2 = O_p(\frac{KL}{\sqrt{N}})$.

$$
\begin{aligned}
E[\|\Omega_3 - \Omega_2\|_2^2] &= E[\|\frac{1}{N}\sum_{i=1}^{N} \rho_i^L(\theta_0)\rho_i^L(\theta_0)' \otimes (q_i^K q_i^{K\prime} - \hat{q}_i^K q_i^{K\prime}\|^2] \\
&= E[\|\frac{1}{N}\sum_{i=1}^{N} A_i \otimes B_i\|^2] \\
&= \frac{1}{N}E[\mathrm{tr}(A_i^2)\mathrm{tr}(B_i^2)] + \frac{N-1}{N}E[(\mathrm{tr}(A_iA_j)\mathrm{tr}(B_iB_j)] \\
&= \frac{1}{N}E[\mathrm{tr}(A_i^2)]E[\mathrm{tr}(B_i^2)] + \frac{N-1}{N}E[(\mathrm{tr}(A_iA_j)]E[\mathrm{tr}(B_iB_j)]
\end{aligned}
$$

Where $A_i = \rho_i^L(\theta_0)\rho_i^L(\theta_0)'$ and $B_i = (q_i^K q_i^{K\prime} - \hat{q}_i^K \hat{q}_i^{K\prime})$, and the last line follows from the independence of A_i and B_i. Note that $E[\mathrm{tr}(A_i^2)] = E[\|\rho_i^L(\theta_0)\|^4] \leq \sqrt{L}^2 L = L^2$, and $E[(\mathrm{tr}(A_iA_j)] = \mathrm{tr}(E[A_i]E[A_j]) = \mathrm{tr}(I_L) = L$. Also:

$$
\begin{aligned}
E[\mathrm{tr}(B_i^2)] &= \mathrm{tr}(E[(q_i^K q_i^{K\prime} - (q_i^K - \bar{q}^K)(q_i^K - \bar{q}^K)')^2]) \\
&= E[\|\bar{q}^K\|^4] \\
&= \frac{1}{N^3} E[\|q_i^K\|^4] + \frac{3(N-1)K}{N^3}
\end{aligned}
$$

Also, since B_i and B_j are not independent, we need to expand the expression, and we get the following:

$$
\begin{aligned}
\mathrm{tr}(E[B_i B_j]) &= \mathrm{tr}(E[(q_i^K q_i^{K\prime} - (q_i^K - \bar{q}^K)(q_i^K - \bar{q}^K)')(q_j^K q_j^{K\prime} - (q_j^K - \bar{q}^K)(q_j^K - \bar{q}^K)')]) \\
&= 4\mathrm{tr}(E[\bar{q}^K(\bar{q}^K)' q_i^K q_j^{K\prime}]) - 3E[\|\bar{q}^K\|^4] \\
&= \frac{8K}{N^2} - \frac{3}{N^3} E[\|q_i^K\|^4] - \frac{9(N-1)K}{N^3}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E[\|\Omega_3 - \Omega_2\|_2^2] &= \frac{1}{N} E[\|\rho_i^L(\theta_0)\|^4](\frac{1}{N^3} E[\|q_i^K\|^4] + \frac{3(N-1)K}{N^3}) + \\
&\quad \frac{N-1}{N} L(\frac{8K}{N^2} - \frac{3}{N^3} E[\|q_i^K\|^4] - \frac{9(N-1)K}{N^3}) \\
&\leq O_p(\frac{\sqrt{L}^2 L \sqrt{K}^2 K}{N^4} + \frac{\sqrt{L}^2 LK}{N^3} + \frac{KL}{N^2} + \frac{L\sqrt{K}^2 K}{N^3}) \\
\Rightarrow \|\Omega_3 - \Omega_2\|_2 &= O_p(\frac{\sqrt{L}\sqrt{K}\sqrt{LK}}{N^2} + \frac{\sqrt{KL}(\sqrt{K} + \sqrt{L})}{N^{3/2}} + \frac{\sqrt{KL}}{N})
\end{aligned}
$$

So, using Markov's inequality, we see that $\|\Omega_3 - \Omega_2\|_2 \xrightarrow{\mathbb{P}} 0$ as long as $\frac{\sqrt{L}\sqrt{K}\sqrt{LK}}{N^2}$, $\frac{\sqrt{KL}(\sqrt{K} + \sqrt{L})}{N^{3/2}}$ and $\sqrt{\frac{KL}{N}}$ all go to 0 as N goes to infinity. Now,

$$
\begin{aligned}
E[\|\Omega_4 - \Omega_3\|_2^2] &= E[\|\frac{1}{N}\sum_{i=1}^{N}(\hat{\rho}_i^L(\theta_0)\hat{\rho}_i^L(\theta_0)' - \rho_i^L(\theta_0)\rho_i^L(\theta_0)') \otimes \hat{q}_i^K\hat{q}_i^{K\prime}\|^2] \\
&= E[\|\frac{1}{N}\sum_{i=1}^{N} A_i \otimes B_i\|^2] \\
&= \frac{1}{N}E[\mathrm{tr}(A_i^2)\mathrm{tr}(B_i^2)] + \frac{N-1}{N}E[(\mathrm{tr}(A_iA_j)\mathrm{tr}(B_iB_j)] \\
&= \frac{1}{N}E[\mathrm{tr}(A_i^2)]E[\mathrm{tr}(B_i^2)] + \frac{N-1}{N}E[(\mathrm{tr}(A_iA_j)]E[\mathrm{tr}(B_iB_j)]
\end{aligned}
$$

Using similarities between these set of expectations and the previous set, we can see that $E[\mathrm{tr}(A_i^2)] = \frac{1}{N^3}E[\|\rho_i^L(\theta_0)\|^4] + \frac{3(N-1)L}{N^3}$ and $E[(\mathrm{tr}(A_iA_j)] = \frac{8L}{N^2} - \frac{3}{N^3}E[\|\rho_i^L(\theta_0)\|^4] - \frac{9(N-1)L}{N^3}$.

Note that $E[\mathrm{tr}(B_i^2)] = E[\|q_i^K\|^4](1 - \frac{4}{N} + \frac{6}{N^2} - \frac{3}{N^3}) + K(N-1)(\frac{6}{N^2} - \frac{9}{N^3})$. Also, $\mathrm{tr}(E[B_iB_j]) = K(1 - \frac{4}{N} + 2\frac{N-1}{N^2} + \frac{8}{N} - 9\frac{N-1}{N^3}) + E[\|q_i^K\|^4](\frac{2}{N^2} - \frac{3}{N^3})$. Adding and multiplying the non-dominated terms, we get that

$$
\begin{aligned}
&E[\|\Omega_4 - \Omega_3\|_2^2] \leq \frac{1}{N^4}(\sqrt{L}^2L + NL)(\sqrt{K}^2K + \frac{K}{N}) + \frac{1}{N^2}(L - \sqrt{L}^2L\frac{1}{N})(K + \frac{\sqrt{K}^2K}{N^2}) \\
&\Rightarrow \|\Omega_4 - \Omega_3\|_2 = O_p(\frac{\sqrt{L}\sqrt{K}\sqrt{LK}}{N^2} + \frac{\sqrt{KL}(\sqrt{K}+\sqrt{L})}{N^{3/2}} + \frac{\sqrt{KL}}{N})
\end{aligned}
$$

Finally, by assumption (1.3.4), we have that $\|\hat{\rho}_i^L(\tilde{\theta}) - \hat{\rho}_i^L(\theta_0)\| \leq \delta_L(W_i)\|\tilde{\theta} - \theta_0\|$ with $\delta_L(W_i) = O_P(\sqrt{L})$, and therefore

$$
\begin{aligned}
\|\Omega_5 - \Omega_4\|_2 &\leq \frac{1}{N}\sum_{i=1}^{N}\|\hat{\rho}_i^L(\tilde{\theta})\hat{\rho}_i^L(\tilde{\theta})' - \hat{\rho}_i^L(\theta_0)\hat{\rho}_i^L(\theta_0)'\|\|\hat{q}_i^K\hat{q}_i^{K\prime}\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}(\|\rho_i^L(\theta_0)\|\|\delta_L(Y_i,W_i)\||\tilde{\theta} - \theta_0| + \|\delta_L(Y_i,W_i)\|^2|\tilde{\theta} - \theta_0|^2)\|q_i^K\|^2
\end{aligned}
$$

but, we have that $\|\rho_i^L(\theta_0)\| = O_p(\sqrt{L})$, $\|\delta_L(Y_i, W_i)\| = O_p(\sqrt{L})$, $|\tilde{\theta} - \theta_0| = O_p(\tau_N)$ and $\|q_i^K\|^2 = O_p(\sqrt{K}^2)$. Combining those, we get that:

$$\begin{aligned}\|\Omega_5 - \Omega_4\| &\leq O_p(\sqrt{K}^2\sqrt{L}\sqrt{L}\tau_N + \sqrt{K}^2 L\tau_N^2) \\ &= O_p(KL\tau_N)\end{aligned}$$

Using the Triangle inequality, we see that if $\|\Omega_{i+1} - \Omega_i\|_2 \xrightarrow{\mathbb{P}} 0$ for $i \in \{1, \ldots, 4\}$, the Hilbert-Schmidt norm of $\tilde{\tilde{\Omega}}^{KL}(\tilde{\theta}) - \Omega^{KL}(\theta_0)$ will converge in probability to 0.

Also, $\|\Omega_5^{-1} - \Omega_1^{-1}\|_2 = \|\Omega_5^{-1}(\Omega_1 - \Omega_5)\Omega_1^{-1}\|_2 \leq \|\Omega_5^{-1}\|_\infty \|\Omega_1 - \Omega_5\|_2 \|\Omega_1^{-1}\|_\infty$. We assumed that the minimum eigenvalue of Ω_1 was uniformly bounded away from 0, therefore $\|\Omega_1^{-1}\|\infty = O_p(1)$. Also, we have that $|\lambda_{\min}(\Omega_5) - \lambda_{\min}(\Omega_1)| \leq \|\Omega_1 - \Omega_5\|_2$, and therefore $\lambda_{\min}(\Omega_5) = O_p(1) + O_p(\|\Omega_1 - \Omega_5\|_2)$, which is also $O_p(1)$ assuming the rate condition is satisfied. Therefore, we also have that $\|\Omega_5^{-1} - \Omega_1^{-1}\|_2 \xrightarrow{\mathbb{P}} 0$.

■

Lemma 1.7.9 *$\|\hat{\tilde{g}}^{KL}(\theta) - E[g^{KL}(\theta)]\| \xrightarrow{\mathbb{P}} 0$ if $\sqrt{\frac{KL}{N}} \to 0$ as $N \to \infty$.*

Proof. Let

$$\begin{aligned}g_1 &= \hat{\tilde{g}}^{KL}(\theta) \\ &= \frac{1}{N}\sum_{i=1}^{N} \hat{\rho}_i^L(\theta) \otimes q_i^K \\ g_2 &= \frac{1}{N}\sum_{i=1}^{N} \rho_i^L(\theta) \otimes q_i^K \\ g_3 &= E[\rho_i^L(\theta) \otimes q_i^K] \\ &= E[g^{KL}(\theta)]\end{aligned}$$

then $\|g_1 - g_3\| \leq \|g_1 - g_2\| + \|g_2 - g_3\|$, and:

$$\begin{aligned}
\|g_1 - g_2\| &= \|\frac{1}{N}\sum_{i=1}^{N}(p_i^L(\theta) - p_i^L(\theta) - (\frac{1}{N}\sum_{j=1}^{N} p_j^L(\theta) - E[p_j^L(\theta)])) \otimes q_i^K\| \\
&= \|\frac{1}{N}\sum_{i=1}^{N}(p_i^L(\theta) - E[p_i^L(\theta)])\|\|\frac{1}{N}\sum_{i=1}^{N} q_i^K\| \\
&= \|\frac{1}{N}\sum_{i=1}^{N}\rho_i^L(\theta)\|\|\frac{1}{N}\sum_{i=1}^{N} q_i^K\| \\
&\leq O_p(\sqrt{\frac{L}{N}})O_p(\sqrt{\frac{K}{N}}) \\
&= O_p(\frac{\sqrt{KL}}{N})
\end{aligned}$$

$$\begin{aligned}
\|g_2 - g_3\| &= \|\frac{1}{N}\sum_{i=1}^{N}(\rho_i^L(\theta) \otimes q_i^K - E[\rho_i^L(\theta) \otimes q_i^K])\| \\
&\leq O_p(\sqrt{\frac{KL}{N}}).
\end{aligned}$$

Therefore, using the triangle inequality, $\|\hat{\tilde{g}}^{KL}(\theta) - E[g^{KL}(\theta)]\| \xrightarrow{\mathbb{P}} 0$. ■

Lemma 1.7.10 *$\hat{R}(\theta) \xrightarrow{\mathbb{P}} R(\theta)$ for all $\theta \in \tilde{\Theta}$ if $K^2L^2(\tau_N + \frac{1}{\sqrt{N}}) \to 0$ as $N \to \infty$.*

Proof. Let

$$
\begin{aligned}
R_1(\theta) &= \hat{R}^{KL}(\theta) \\
&= \hat{\bar{g}}^{KL}(\theta)' \hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1} \hat{\bar{g}}^{KL}(\theta) \\
R_2(\theta) &= \hat{\bar{g}}^{KL}(\theta)' \Omega^{KL}(\theta_0)^{-1} \hat{\bar{g}}^{KL}(\theta) \\
R_3(\theta) &= E[g^{KL}(\theta)]' \Omega^{KL}(\theta_0)^{-1} E[g^{KL}(\theta)] \\
R_4(\theta) &= R(\theta)
\end{aligned}
$$

$$
\begin{aligned}
|R_1(\theta) - R_2(\theta)| &= |\hat{\bar{g}}^{KL}(\theta)' (\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1} - \Omega^{KL}(\theta_0)^{-1}) \hat{\bar{g}}^{KL}(\theta)| \\
&\leq \|\hat{\bar{g}}^{KL}(\theta)\|^2 \|\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1} - \Omega^{KL}(\theta_0)^{-1}\| \\
&= (\|g^{KL}(\theta)\| + \|\hat{\bar{g}}^{KL}(\theta) - E[g^{KL}(\theta)]\|)^2 \|\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1} - \Omega^{KL}(\theta_0)^{-1}\| \\
&= (O_p(\sqrt{K}\sqrt{L}) + O_p(\sqrt{\frac{KL}{N}}))^2 \times \\
& O_p(KL\tau_N + \sqrt{\frac{KL}{N}}\sqrt{K}\sqrt{L} + \sqrt{\frac{KL}{N}}\frac{\sqrt{K}+\sqrt{L}}{N})
\end{aligned}
$$

also, we have that:

$$
\begin{aligned}
|R_2(\theta) - R_3(\theta)| &\leq |(\hat{\bar{g}}^{KL}(\theta) - E[g^{KL}(\theta)])' \Omega^{KL}(\theta_0)^{-1} (\hat{\bar{g}}^{KL}(\theta) - E[g^{KL}(\theta)])| \\
&\leq \|\hat{\bar{g}}^{KL}(\theta) - E[g^{KL}(\theta)]\|^2 \|\Omega^{KL}\| \\
&= O_p(\sqrt{\frac{KL}{N}})^2 O_p(\sqrt{KL}) \\
&= O_p(\frac{K^{3/2}L^{3/2}}{N})
\end{aligned}
$$

Finally,

$$
\begin{aligned}
|R_3(\theta) - R_4(\theta)| &= |\|E[g^{KL}(\theta)]\|^2_{\Omega^{KL}} - \|E[g(\cdot,\cdot,\theta)]\|^2_{\Omega}| \\
&\xrightarrow{\mathbb{P}} 0
\end{aligned}
$$

where $\|g\|^2_{\Omega^{KL}} = g'\Omega^{KL^{-1}}g$, and if K and $L \to \infty$. The proof of this fact can be found in Parzen (1959), and relies on assumption (1.3.1)(c). Using the triangle inequality, we have shown that $\hat{R}^{KL}(\theta) \xrightarrow{\mathbb{P}} R(\theta)$ for all $\theta \in \tilde{\Theta}$. ■

Proof of Theorem (1.3.5)**.** Using lemmas (1.7.7), (1.7.6),(1.7.10) and assumption (1.3.4) we have the following:

- Compact parameter space $\tilde{\Theta}$
- $R(\theta)$ is continuous on $\tilde{\Theta}$
- $|\hat{R}^{KL}(\theta) - \hat{R}^{KL}(\theta')| \leq \hat{D}|\theta - \theta'|^{\alpha}$ with $\hat{D} = O_p(1)$
- $\hat{R}^{KL}(\theta) \xrightarrow{\mathbb{P}} R(\theta)$ for all $\theta \in \tilde{\Theta}$

By Newey and McFadden (1994), 3 and 4 imply uniform convergence of $\hat{R}^{KL}(\theta)$ to $R(\theta)$ which then implies consistency of the estimator. ■

1.7.3 Asymptotic Normality under Unconditional Independence

Let $\hat{\bar{G}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \hat{\rho}_{i\theta}^{L}(\theta) \otimes \hat{q}_i^K$, $\bar{G}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \rho_{i\theta}^{L}(\theta) \otimes q_i^K$ and $G(\theta) = E[\rho_{i\theta}^{L}(\theta) \otimes q_i^K]$. Also,

Lemma 1.7.11 $\|\hat{\bar{G}}(\tilde{\theta}) - E[G(\theta_0)]\| \xrightarrow{\mathbb{P}} 0$ *if* $\tau_N\sqrt{KL} + \sqrt{\frac{KL}{N}} \to 0$ *as* $N \to \infty$, *where* $\|\tilde{\theta} - \theta_0\| = O_p(\tau_N)$.

Proof.

$$\begin{aligned}
\|\hat{\bar{G}}(\tilde{\theta}) - \hat{\bar{G}}(\theta_0)\| &= \|\frac{1}{N}\sum_{i=1}^{N}(\hat{\rho}_{i\theta}^{L}(\tilde{\theta}) - \hat{\rho}_{i\theta}^{L}(\theta)) \otimes (e^{i\vec{t}_L X_i} - \frac{1}{N}\sum_{j=1}^{N} e^{i\vec{t}_L X_j})\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \|\hat{\rho}_{i\theta}^{L}(\tilde{\theta}) - \hat{\rho}_{i\theta}^{L}(\theta)\| \|e^{i\vec{t}_L X_i} - \frac{1}{N}\sum_{j=1}^{N} e^{i\vec{t}_L X_j}\| \\
&\leq |\tilde{\theta} - \theta_0| \frac{1}{N}\sum_{i=1}^{N} \tilde{\delta}_L(W_i) \|e^{i\vec{t}_L X_i} - \frac{1}{N}\sum_{j=1}^{N} e^{i\vec{t}_L X_j}\| \\
&= O_p(\tau_N\sqrt{LK}).
\end{aligned}$$

Also, by arguments similar to those in (1.7.9), we will have that

$$\|\hat{\bar{G}}(\theta_0) - \bar{G}(\theta_0)\| = O_p(\frac{\sqrt{K}\sqrt{L}}{\sqrt{N}})$$
$$\|\bar{G}(\theta_0) - E[G(\theta_0)]\| = O_p(\frac{\sqrt{K}\sqrt{L}}{\sqrt{N}})$$

Using the triangle inequality, we see that the rate condition stated in the lemma is sufficient to ensure that $\|\hat{\tilde{G}}(\tilde{\theta}) - E[G(\theta_0)]\| \xrightarrow{\mathbb{P}} 0$. ■

Lemma 1.7.12 *The efficiency bound from the continuum of moment conditions is equal to that computed using the projection method.* $\|\frac{\partial}{\partial\theta}g(\theta_0)\|_\Omega^{-2} = V_{eff}^{-1}(\theta_0)$

Proof.

$$\begin{aligned}
\|\frac{\partial}{\partial\theta}g(\theta_0)\|_\Omega^2 &= (\frac{\partial}{\partial\theta}g(\theta_0), \Omega^{-1}\frac{\partial}{\partial\theta}g(\theta_0)) \\
\frac{\partial}{\partial\theta}g(s,t,\theta_0) &= \text{Cov}\left(is\rho_\theta(Y,W,\theta_0)e^{is\epsilon}, e^{itX}\right) \\
&= E[ise^{is\epsilon}W(X,\epsilon)e^{itX}] - E[ise^{is\epsilon}W(X,\epsilon)]E[e^{itX}]
\end{aligned}$$

where $W(X,\epsilon) = E[\rho_\theta(W,\theta_0)|X,\epsilon]$.

$$\begin{aligned}
&= E[ise^{is\epsilon}W(X,\epsilon)e^{itX}] - E[ise^{is\epsilon}W(X,\epsilon)]E[e^{itX}] \\
&= -\int\int \frac{W_\epsilon(X,\epsilon)f_\epsilon(\epsilon) + W(X,\epsilon)f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}e^{itX}e^{is\epsilon}d\epsilon dX + \\
&\int\int \frac{W_\epsilon(X,\epsilon)f_\epsilon(\epsilon) + W(X,\epsilon)f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}e^{is\epsilon}d\epsilon E[e^{itX}] \\
&= -(E[\Lambda(X,\epsilon)e^{itX}e^{it\epsilon}] - E[\Lambda(X,\epsilon)e^{it\epsilon}]E[e^{itX}]
\end{aligned}$$

where $\Lambda(X,\epsilon) = \dfrac{W_\epsilon(X,\epsilon)f_\epsilon(\epsilon) + W(X,\epsilon)f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}$.
Using the Fourier transform, we can write:

$$\Lambda(X,\epsilon) = \int\int e^{it'X+is'\epsilon}\tilde{\Lambda}(s',t')ds'dt'$$

where $\tilde{\Lambda}$ is the Fourier transform of Λ. Therefore,

$$
\begin{aligned}
&= -(E[\Lambda(X,\epsilon)e^{itX}e^{it\epsilon}] - E[\Lambda(X,\epsilon)e^{it\epsilon}]E[e^{itX}] \\
&= -\int\int\int\int e^{i(s+s')\epsilon}e^{i(t+t')X}d\epsilon dX\tilde{\Lambda}(s',t')ds'dt' + \\
&\int\int\int\int e^{i(s+s')\epsilon}e^{it'X}d\epsilon dX\tilde{\Lambda}(s',t')ds'dt'E[e^{itX}] \\
&= -\Omega\tilde{\Lambda}(s,t)
\end{aligned}
$$

Therefore,
$\|\frac{\partial}{\partial\theta}g(\theta_0)\|^2_\Omega = (\Omega\tilde{\Lambda}, \Omega^{-1}\Omega\tilde{\Lambda}) = (\Omega\tilde{\Lambda}, \tilde{\Lambda})$, and then:

$$
\begin{aligned}
(\Omega\tilde{\Lambda}, \tilde{\Lambda}) &= \int\int \Omega\tilde{\Lambda}(s,t)\tilde{\Lambda}(s,t)dsdt \\
&= \int\int\int\int k(s',t',s,t)\tilde{\Lambda}(s',t')\tilde{\Lambda}(s,t)dsds'dtdt' \\
&= \int\cdots\int e^{i(s+s')\epsilon}e^{i(t+t')X}\tilde{\Lambda}(s',t')\tilde{\Lambda}(s,t)f_X(X)f_\epsilon(\epsilon)dsds'dtdt'dXd\epsilon - \\
&\int\cdots\int e^{i(s+s')\epsilon}e^{itX+it'X'}\tilde{\Lambda}(s',t')\tilde{\Lambda}(s,t)f_X(X)f_\epsilon(\epsilon)dsds'dtdt'dXdX'd\epsilon \\
&= \int\int \Lambda(X,\epsilon)^2 f_X(X)f_\epsilon(\epsilon)dXd\epsilon - \\
&\int\int\int \Lambda(X,\epsilon)\Lambda(X',\epsilon)f_X(X)f_X(X')f_\epsilon(\epsilon)dXdX'd\epsilon \\
&= E[\Lambda(X,\epsilon)^2] - E[E[\Lambda(X,\epsilon)|\epsilon]^2] \\
&= V_{\text{eff}}^{-1}
\end{aligned}
$$

■

Lemma 1.7.13 *If,* $K^2L^2(\tau_N + \frac{1}{\sqrt{N}}) \to 0$, *then* $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V_{\text{eff}})$.

Proof. We have that $\sqrt{N}(\hat{\theta} - \theta_0) = -\sqrt{N}\frac{\partial}{\partial\theta}\hat{R}^{KL}(\theta_0)(\frac{\partial^2}{\partial\theta^2}\hat{R}^{KL}(\bar{\theta}))^{-1}$, with $\bar{\theta}$ in between θ_0 and $\hat{\theta}$. We can see that:

$$\frac{\partial^2}{\partial\theta^2}\hat{R}^{KL}(\bar{\theta}) = 2\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) + 2\hat{\bar{G}}_\theta(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\bar{\theta})$$

$$\begin{aligned}
|\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) - \|E[G(\cdot,\theta_0)]\|_\Omega^2| &\leq |\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) - \hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KL}(\theta_0)^{-1}\hat{\bar{G}}(\theta_0)| \\
&\quad + |\hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KL}(\theta_0)^{-1}\hat{\bar{G}}(\theta_0) - \hat{\bar{G}}(\theta_0)'\Omega^{KL}(\theta_0)^{-1}\hat{\bar{G}}(\theta_0)| \\
&\quad + |\hat{\bar{G}}(\theta_0)'\Omega^{KL}(\theta_0)^{-1}\hat{\bar{G}}(\theta_0) - G(\theta_0)'\Omega^{KL}(\theta_0)^{-1}G(\theta_0)| \\
&\quad + |G(\theta_0)'\Omega^{KL}(\theta_0)^{-1}G(\theta_0) - \|E[G(\cdot,\theta_0)]\|_\Omega^2| \\
&= (1)+(2)+(3)+(4) \\
(1) &\leq 2|\hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) + \\
&\quad (\hat{\bar{G}}(\bar{\theta}) - \hat{\bar{G}}(\theta_0))'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}(\hat{\bar{G}}(\bar{\theta}) - \hat{\bar{G}}(\theta_0))| \\
&\leq O_p(\sqrt{N}KL) + O_p(NKL) \\
&= O_p(\sqrt{N}KL) \\
(2) &\leq O_p(\tau_N K^2 L^2) \\
(3) &\leq O_p(\frac{KL}{N}) \\
(4) &= o_P(1)
\end{aligned}$$

$$\begin{aligned}
\hat{\bar{G}}_\theta(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\bar{\theta}) &\leq \|\hat{\bar{G}}_\theta(\bar{\theta})\|\|\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\|_\infty\|\hat{\bar{g}}(\bar{\theta})\| \\
&\leq O_p(\sqrt{KL})O_p(1)O_p(\tau_N\sqrt{KL})\|
\end{aligned}$$

Also,

$$\begin{aligned}
\sqrt{N}\frac{\partial}{\partial\theta}\hat{R}^{KL}(\theta_0) = 2\hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0)\|\sqrt{N}\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0) - \\
\sqrt{N}E[G^{KL}(\theta_0)]'\Omega^{KL}(\theta_0)^{-1}\bar{g}(\theta_0)\| \\
\xrightarrow{\mathbb{P}} 0
\end{aligned}$$

since,

$$\begin{aligned}\sqrt{N}\|\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0) - E[G^{KL}(\theta_0)]'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0)\| \leq \\ \sqrt{N}\|\hat{\bar{G}}(\theta_0) - E[G^{KL}(\theta_0)]\|\|\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\|\|\hat{\bar{g}}(\theta_0)\| \\ = \sqrt{N}O_p(\tau_N\sqrt{KL} + \sqrt{\frac{KL}{N}})O_p(KL)O_p(\tau_N\sqrt{KL} + \sqrt{\frac{KL}{N}}) \\ \xrightarrow{\mathbb{P}} 0\end{aligned}$$

$$\begin{aligned}\sqrt{N}\|E[G^{KL}(\theta_0)]'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0) - E[G^{KL}(\theta_0)]'\Omega^{KL}(\theta_0)^{-1}\bar{g}(\theta_0)\| \leq \sqrt{N}\|E[G^{KL}(\theta_0)]\| \\ \cdot\|\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0) - \Omega^{KL}(\theta_0)^{-1}\bar{g}(\theta_0)\| \\ = \sqrt{N}O_p(\sqrt{KL})O_p(\frac{K^{3/2}L^{3/2}\tau_N}{\sqrt{N}}) \\ \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

If these conditions are satisfied, $\sqrt{N}E[G^{KL}(\theta_0)]'\Omega^{KL}(\theta_0)^{-1}\bar{g}(\theta_0)$ converges by a standard CLT to a $N(0, V_{\text{eff}}^{-1})$ random variable. Combining these two facts, we get the result. ■

1.7.4 Consistency under Conditional Independence

Let $g(s,t,u,\theta) = e^{is\rho(Y,W,\theta)}(e^{itX} - E[e^{itX}|Z])e^{iuZ}$, and let Ω be the operator defined by the kernel $k(s,t,u,s',t',u') = E[\text{Cov}\left(e^{is\epsilon}, e^{is'\epsilon}|Z\right)\text{Cov}\left(e^{itX}, e^{it'X}|Z\right)e^{i(u+u')Z}]$. Let $R(\theta) = \|E[g(\cdot,\cdot,\cdot,\theta)]\|_\Omega^2$.

Lemma 1.7.14 (Identification) $R(\theta) = 0 \Rightarrow \theta = \theta_0$

Proof.

$$\begin{aligned}R(\theta) = 0 &\Rightarrow \Omega^{-1/2}E[g(\cdot,\cdot,\cdot\theta)] = 0 \\ &\Leftrightarrow E[g(\cdot,\cdot,\cdot,\theta)] = 0 \text{ Since } \Omega^{-1/2} \text{ is an invertible operator} \\ &\Leftrightarrow \text{Cov}\left(e^{is\rho(Y,W,\theta)}, e^{itX}|Z\right) = 0 \text{ for all } s,t \in \mathbb{R} \\ &\Leftrightarrow \rho(Y,W,\theta) \perp X|Z \text{ :complex exponentials are a convergence determining class} \\ &\Leftrightarrow \theta = \theta_0 \text{ from the identifying assumption.}\end{aligned}$$

■

Lemma 1.7.15 (Continuity) *$R(\theta)$ is continuous in θ for all $\theta \in \tilde{\Theta}$, a compact subset of the parameter space Θ.*

Proof.

Proof is similar to that of (1.7.7), and uses the smoothness in $\rho(\cdot,\cdot,\theta)$ and the conditional density of ϵ given Z ■

Let $p_i^L(\theta)$ denote the basis for $\rho(Y,W,\theta)$, q_i^K the basis for X and t_i^M the basis for W. Let $\lambda(K,Z) = E[q_i^K|Z]$, and $\widehat{\lambda}(K,Z)$ be its estimate. Let

$$\widehat{\bar{\Omega}}^{KLM}(\theta) = \frac{1}{N}\sum_{i=1}^{N} p_i^L(\theta)p_i^L(\theta)' \otimes (q_i^K - \widehat{\lambda}(K,Z))(q_i^K - \widehat{\lambda}(K,Z))' \otimes t_i^M t_i^{M\prime},$$

$$\bar{\Omega}^{KLM}(\theta) = \frac{1}{N}\sum_{i=1}^{N} p_i^L(\theta)p_i^L(\theta)' \otimes (q_i^K - \lambda(K,Z))(q_i^K - \lambda(K,W))' \otimes t_i^M t_i^{M\prime},$$

$$\Omega^{KLM}(\theta) = E[p_i^L(\theta)p_i^L(\theta)' \otimes (q_i^K - \lambda(K,Z))(q_i^K - \lambda(K,Z))' \otimes t_i^M t_i^{M\prime}].$$

Lemma 1.7.16 *Let $\tilde{\theta} - \theta_0 = O_p(\tau_N)$, and $\|\widehat{\lambda}(K,Z) - \lambda(K,Z)\| = O_p(\nu_N)$, and then $\|\widehat{\bar{\Omega}}^{KLM}(\tilde{\theta}) - \Omega^{KLM}(\theta_0)\|_2 \xrightarrow{\mathbb{P}} 0$ if $KLM\tau_N \to 0$, $\sqrt{\frac{KLM}{N}}\sqrt{K}\sqrt{L}\zeta(M) \to 0$, $LM\sqrt{K}\nu_N + LM\nu_N^2$ and$\sqrt{\frac{KLM}{N}}\frac{\sqrt{K}+\sqrt{L}+\zeta(M)}{N} \to 0$ Also, $\|\widehat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1} - \Omega^{KLM}(\theta_0)^{-1}\|_2 \xrightarrow{\mathbb{P}} 0$ if the same rates for K and L are chosen.*

Proof. Let

$$\Omega_1 = \Omega^{KLM}(\theta_0)$$
$$\Omega_2 = \bar{\Omega}^{KLM}(\theta_0)$$
$$\Omega_3 = \bar{\Omega}^{KLM}(\tilde{\theta})$$
$$\Omega_4 = \widehat{\bar{\Omega}}^{KLM}(\tilde{\theta})$$

By arguments similar to those in the proof of (1.7.8), we will have that $\|\Omega_3 - \Omega_1\|_2 \xrightarrow{\mathbb{P}} 0$ when $KLM\tau_N \to 0$, $\sqrt{\frac{KLM}{N}}\sqrt{K}\sqrt{L}\zeta(M) \to 0$ and$\sqrt{\frac{KLM}{N}}\frac{\sqrt{K}+\sqrt{L}+\zeta(M)}{N} \to 0$. Let us now conisder $\|\Omega_4 - \Omega_3\|_2$:

$$\begin{aligned}\|\Omega_4 - \Omega_3\|_2 \leq \frac{1}{N}\sum_{i=1}^{N} \|p_i^L(\tilde{\theta})p_i^L(\tilde{\theta})'\|_2 \|(q_i^K - \widehat{\lambda}(K,Z))(q_i^K - \widehat{\lambda}(K,Z))' - \\ (q_i^K - \lambda(K,Z))(q_i^K - \lambda(K,Z))'\|_2 \times \\ \|t_i^M t_i^{M\prime}\|_2 \\ \leq O_p(LM)\frac{1}{N}\sum_{i=1}^{N}(2\|q_i^K(\widehat{\lambda}(K,Z) - \lambda(K,Z))'\|_2 + \\ 2\|\lambda(K,Z)(\widehat{\lambda}(K,Z) - \lambda(K,Z))'\|_2 + \|(\widehat{\lambda}(K,Z) - \lambda(K,Z))(\widehat{\lambda}(K,Z) - \lambda(K,Z))'\|_2) \\ \leq O_p(LM)\frac{1}{N}\sum_{i=1}^{N}(4\sqrt{K}\nu_N + \nu_N^2) \\ = O_p(LM\sqrt{K}\nu_N + LM\nu_N^2)\end{aligned}$$

■

Lemma 1.7.17 $\|\hat{\bar{g}}^{KLM}(\theta) - E[g^{KL}(\theta)]\| \xrightarrow{\mathbb{P}} 0$ *if* $\sqrt{\frac{KLM}{N}} \to 0$ *and* $\sqrt{LM}\nu_N \to 0$ *as* $N \to \infty$.

Proof. Let

$$\begin{aligned}g_1 &= \hat{\bar{g}}^{KLM}(\theta) \\ &= \frac{1}{N}\sum_{i=1}^{N} p_i^L(\theta) \otimes (q_i^K - \widehat{\lambda}(K,Z)) \otimes t_i^M \\ g_2 &= \frac{1}{N}\sum_{i=1}^{N} p_i^L(\theta) \otimes (q_i^K - \lambda(K,Z)) \otimes t_i^M \\ g_3 &= E[\rho_i^L(\theta) \otimes (q_i^K - \lambda(K,Z)) \otimes t_i^M] \\ &= E[g^{KLM}(\theta)]\end{aligned}$$

then $\|g_1 - g_3\| \leq \|g_1 - g_2\| + \|g_2 - g_3\|$, and:

$$\|g_1 - g_2\| = \|\frac{1}{N}\sum_{i=1}^{N} p_i^L(\theta) \otimes (\lambda(K,Z) - \widehat{\lambda}(K,Z)) \otimes t_i^M\|$$
$$\leq \frac{1}{N}\sum_{i=1}^{N} \|p_i^L(\theta)\|\|\lambda(K,Z) - \widehat{\lambda}(K,Z)\|\|t_i^M\|$$
$$= O_p(\sqrt{KM}\nu_N)$$

$$\|g_2 - g_3\| \leq O_p(\sqrt{\frac{KLM}{N}}).$$

Therefore, using the triangle inequality, $\|\hat{\tilde{g}}^{KLM}(\theta) - E[g^{KLM}(\theta)]\| \xrightarrow{\mathbb{P}} 0$. ■

Lemma 1.7.18 *$\hat{R}(\theta) \xrightarrow{\mathbb{P}} R(\theta)$ for all $\theta \in \tilde{\Theta}$ if $K^2L^2M^2(\tau_N + \frac{1}{\sqrt{N}}) \to 0$ and $(\sqrt{K}\sqrt{L}\zeta(M))^2 LM\nu_N^2 \to 0$ as $N \to \infty$.*

Proof. The proof uses lemmas (1.7.17) and (1.7.16), and is analogous to that of lemma (1.7.10). ■

Proof of Theorem (1.3.7)**.** Using lemmas (1.7.15), (1.7.14),(1.7.18) and assumptions (1.3.6) we have the following:

- Compact parameter space $\tilde{\Theta}$
- $R(\theta)$ is continuous on $\tilde{\Theta}$
- $|\hat{R}^{KLM}(\theta) - \hat{R}^{KLM}(\theta')| \leq \hat{D}|\theta - \theta'|^{\alpha}$ with $\hat{D} = O_p(1)$
- $\hat{R}^{KLM}(\theta) \xrightarrow{\mathbb{P}} R(\theta)$ for all $\theta \in \tilde{\Theta}$

By Newey and McFadden (1994), 3 and 4 imply uniform convergence of $\hat{R}^{KLM}(\theta)$ to $R(\theta)$ which then implies consistency of the estimator. ■

1.7.5 Asymptotic normality under Conditional Independence

Let $\hat{\bar{G}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} p_{i\theta}^L(\theta) \otimes (q_i^K - \widehat{\lambda}(K,W)) \otimes t_i^M$, $\bar{G}(\theta) = \frac{1}{N}\sum_{i=1}^{N} p_{i\theta}^L(\theta) \otimes (q_i^K - \lambda(K,W)) \otimes t_i^M$ and $G(\theta) = E[p_{i\theta}^L(\theta) \otimes (q_i^K - \lambda(K,W)) \otimes t_i^M]$.

Lemma 1.7.19 *$\|\hat{\bar{G}}(\tilde{\theta}) - E[G(\theta_0)]\| \xrightarrow{\mathbb{P}} 0$ if $\tau_N\sqrt{KLM} + \sqrt{\frac{KLM}{N}} + LM\nu_N \to 0$ as $N \to \infty$, where $\|\tilde{\theta} - \theta_0\| = O_p(\tau_N)$.*

Proof. The proof is analogous to (1.7.11). ■

Lemma 1.7.20 *The efficiency bound from the continuum of moment conditions is equal to that computed using the projection method.* $\|\frac{\partial}{\partial\theta}g(\theta_0)\|_\Omega^{-2} = V_{eff}^{-1}(\theta_0)$

Proof.

$$\begin{aligned}
\|\frac{\partial}{\partial\theta}g(\theta_0)\|_\Omega^2 &= (\frac{\partial}{\partial\theta}g(\theta_0), \Omega^{-1}\frac{\partial}{\partial\theta}g(\theta_0)) \\
\frac{\partial}{\partial\theta}g(s,t,u,\theta_0) &= E[\text{Cov}\left(is\rho_\theta(Y,W,\theta_0)e^{is\epsilon}, e^{itX}|Z\right)e^{iuZ}] \\
&= E[ise^{is\epsilon}V(X,\epsilon,Z)e^{itX}e^{iuZ}] - E[E[ise^{is\epsilon}V(X,\epsilon,Z)|Z]E[e^{itX}|Z]e^{iuZ}]
\end{aligned}$$

where $V(X,\epsilon,W) = E[\rho_\theta(Y,W,\theta_0)|X,\epsilon,Z]$.

$$\begin{aligned}
&= E[ise^{is\epsilon}V(X,\epsilon,Z)e^{itX}e^{iuZ}] - E[E[ise^{is\epsilon}V(X,\epsilon,Z)|Z]E[e^{itX}|Z]e^{iuZ}] \\
&= -(E[\Lambda(X,\epsilon,Z)e^{itX}e^{it\epsilon}e^{iuZ}] - E[E[\Lambda(X,\epsilon,Z)e^{it\epsilon}|Z]E[e^{itX}|Z]e^{iuZ}]
\end{aligned}$$

where $\Lambda(X,\epsilon,Z) = \dfrac{V_\epsilon(X,\epsilon,Z)f_{\epsilon|Z}(\epsilon|Z) + V(X,\epsilon,Z)f'_{\epsilon|Z}(\epsilon|Z)}{f_{\epsilon|Z}(\epsilon|Z)}$.
Using the Fourier transform, we can write:

$$\Lambda(X,\epsilon,W) = \int\int\int e^{it'X+is'\epsilon+iu'W}\tilde{\Lambda}(s',t',u')ds'dt'du'$$

where $\tilde{\Lambda}$ is the Fourier transform of Λ. Therefore,

$$\begin{aligned}
&= -(E[\Lambda(X,\epsilon,Z)e^{itX}e^{it\epsilon}e^{iuZ}] - E[E[\Lambda(X,\epsilon,Z)e^{it\epsilon}|Z]E[e^{itX}|Z]e^{iuZ}] \\
&= -\int\cdots\int e^{i(s+s')\epsilon}e^{i(t+t')X}e^{i(u+u')Z}f(\epsilon,X,Z)d\epsilon dXdZ\tilde{\Lambda}(s',t',u')ds'dt'du' \\
&+ \int\cdots\int e^{i(s+s')\epsilon}e^{it'X}f(\epsilon|Z)d\epsilon f(X|Z)dXE[e^{itX}|Z]e^{i(u+u')Z}f(Z)dZ\tilde{\Lambda}(s',t',u')ds'dt'du' \\
&= -\Omega\tilde{\Lambda}(s,t,u)
\end{aligned}$$

Therefore,

$\|\frac{\partial}{\partial\theta}g(\theta_0)\|^2_\Omega = (\Omega\tilde{\Lambda}, \Omega^{-1}\Omega\tilde{\Lambda}) = (\Omega\tilde{\Lambda}, \tilde{\Lambda})$, and then:

$$\begin{aligned}
(\Omega\tilde{\Lambda}, \tilde{\Lambda}) &= \int\int\int \Omega\tilde{\Lambda}(s,t,u)\tilde{\Lambda}(s,t)dsdtdu \\
&= \int\int\int\int\int\int k(s',t',u',s,t,u)\tilde{\Lambda}(s',t',u')\tilde{\Lambda}(s,t,u)dsds'dtdt'dudu' \\
&= E[\Lambda(X,\epsilon,Z)^2] - E[E[\Lambda(X,\epsilon,Z)|\epsilon,Z]^2] \\
&= V_{\text{eff}}^{-1}
\end{aligned}$$

■

Lemma 1.7.21 *If $K^2L^2M^2(\tau_N + \frac{1}{\sqrt{N}}) \to 0$ as $N \to \infty$, then $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V_{\text{eff}})$.*

Proof. We have that $\sqrt{N}(\hat{\theta} - \theta_0) = -\sqrt{N}\frac{\partial}{\partial\theta}\hat{R}^{KLM}(\theta_0)(\frac{\partial^2}{\partial\theta^2}\hat{R}^{KLM}(\bar{\theta}))^{-1}$, with $\bar{\theta}$ lying between θ_0 and $\hat{\theta}$. We can see that:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\hat{R}^{KLM}(\bar{\theta}) &= 2\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) + 2\hat{\bar{G}}_\theta(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\bar{\theta}) \\
|\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) - \|E[G(\cdot,\theta_0)]\|^2_\Omega| &\leq |\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{G}}(\bar{\theta}) - \hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{G}}(\theta_0)| \\
&\quad + |\hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KL}(\tilde{\theta})^{-1}\hat{\bar{G}}(\theta_0) - \hat{\bar{G}}(\theta_0)'\Omega^{KLM}(\theta_0)^{-1}\hat{\bar{G}}(\theta_0)| \\
&\quad + |\hat{\bar{G}}(\theta_0)'\Omega^{KLM}(\theta_0)^{-1}\hat{\bar{G}}(\theta_0) - G(\theta_0)'\Omega^{KLM}(\theta_0)^{-1}G(\theta_0)| \\
&\quad + |G(\theta_0)'\Omega^{KLM}(\theta_0)^{-1}G(\theta_0) - \|E[G(\cdot,\theta_0)]\|^2_\Omega| \\
&= (1) + (2) + (3) + (4) \\
(1) &\leq O_p(\frac{KLM}{\sqrt{N}}) + O_p(\frac{KLM}{N}) \\
&= O_p(\frac{KLM}{\sqrt{N}}) \\
(2) &\leq O_p(K^2L^2M^2\tau_N + KL^2M^2\nu_N^2) \\
(3) &\leq O_p(\frac{KLM}{N}) \\
(4) &= o_P(1)
\end{aligned}$$

$$\begin{aligned}
\hat{\bar{G}}_\theta(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\bar{\theta}) &\leq \|\hat{\bar{G}}_\theta(\bar{\theta})\|\|\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\|_\infty\|\hat{\bar{g}}(\bar{\theta})\| \\
&\leq O_p(\sqrt{KLM})O_p(1)O_p(\tau_N\sqrt{KLM} + \nu_N\sqrt{LM})\|
\end{aligned}$$

Also,

$$\begin{aligned}\sqrt{N}\frac{\partial}{\partial\theta}\hat{R}^{KLM}(\theta_0) = 2\hat{\bar{G}}(\theta_0)'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0)\|\sqrt{N}\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0)- \\ \sqrt{N}E[G(\theta_0)]'\Omega^{KLM}(\theta_0)^{-1}\bar{g}(\theta_0)\| \\ \xrightarrow{\mathbb{P}} 0\end{aligned}$$

since,

$$\begin{aligned}\sqrt{N}\|\hat{\bar{G}}(\bar{\theta})'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0) - E[G(\theta_0)]'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0)\| \leq \\ \sqrt{N}\|\hat{\bar{G}}(\theta_0) - E[G(\theta_0)]\|\|\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\|_\infty\|\hat{\bar{g}}(\theta_0)\| \\ = \sqrt{N}O_p(\tau_N\sqrt{KLM} + \sqrt{\frac{KLM}{N}} + \sqrt{LM}\nu_N)\times \\ O_p(1)O_p(\tau_N\sqrt{KLM} + \sqrt{\frac{KLM}{N}} + \sqrt{LM}\nu_N) \\ \xrightarrow{\mathbb{P}} 0\end{aligned}$$

$$\begin{aligned}\sqrt{N}\|E[G(\theta_0)]'\hat{\bar{\Omega}}^{KLM}(\tilde{\theta})^{-1}\hat{\bar{g}}(\theta_0) - E[G^{KL}(\theta_0)]'\Omega^{KLM}(\theta_0)^{-1}\bar{g}(\theta_0)\| \leq \\ \sqrt{N}O_p(\sqrt{KLM})O_p\left(\frac{K^{3/2}L^{3/2}M^{3/2}\tau_N + K^{1/2}L^{3/2}M^{3/2}\nu_N}{\sqrt{N}}\right) \\ \xrightarrow{\mathbb{P}} 0\end{aligned}$$

if these conditions are satisfied, $\sqrt{N}E[G(\theta_0)]'\Omega^{KLM}(\theta_0)^{-1}\bar{g}(\theta_0)$ converges by a standard CLT to a $N(0, V_{\text{eff}}^{-1})$ random variable. Combining these two facts, we get the result.

■

Chapter 2

Estimation of Quantile Effects in Panel Data with Random Coefficients

2.1 Introduction

The recent paper Graham and Powell (2012) introduced a panel data model with correlated random coefficients. The panel structure allows to identify individual effects through observing repeated observations for each unit. The coefficients in their model are random, and might depend on the covariates. This relaxes the classical fixed coefficients approach, which imposes that treatment effects are equal across different units. The goal of that paper was to identify and estimate the average partial effect (APE), which is the average (over the covariates) of the random coefficient. This measure is related to the average structural function and can be interpreted as the average of the (random) effect of an increase by 1 of the covariates. Their results are derived under a continuous covariates assumption, and just identification, where the number of time periods equals the number of random coefficients. Under these assumptions, the APE estimator converges at a rate slower than root-N due to the irregular identification, in contrast to Chamberlain (1982) which covers the over-identified case.

This chapter studies a model similar to that in Graham and Powell (2012) but focuses on the estimation of quantiles of the distribution of the correlated random coefficients. We define the ACQE, the average conditional quantile effect, and the UQE, the unconditional quantile effect as a function of the distribution of the random coefficient. The ACQE is the average (over the covariates) of the distribution of the coefficients, conditional on the covariates. It is a measure of the average magnitude of the random coefficient. The UQE is the unconditional quantile of the distribution of the random coefficients. For example, we can interpret the median UQE as the median size of the effect of an exogenous increase in a covariate.

In this chapter, we derive estimators for the ACQE and the UQE in a just-identified panel data model with discrete covariates. The first case we consider, that we term the regular case, has covariates distributed discretely with a mass point of stayers, i.e. units with determinant equal to 0. We show the root-N consistency of our estimators and compute their asymptotic distribution. The second case we consider, the bandwidth case, has a shrinking mass of

"near-stayers" with determinant shrinking to 0, and a shrinking mass of exact stayers. We propose this model to approximate a model where covariates are distributed continuously, as is the case in Graham and Powell (2012). We show the root-Nh_N consistency of estimators for the ACQE and UQE, compute their asymptotic distribution and then draw parallels to the continuous case.

We introduce the model in the next section, then present additional assumptions for the two different cases proposed here. We then show our asymptotic results for the ACQE and the UQE, before concluding.

2.2 General Model

Let $\mathbf{Y} = (Y_1, \ldots, Y_T)'$ be a $T \times 1$ vector of outcomes and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)'$ a $T \times P$ matrix of regressors with $\mathbf{X}_t \in \mathbb{X}_{tN} \subset \mathbb{R}^P$ and $\mathbf{X} \in \mathbb{X}_N^T$ where $\mathbb{X}_N^T = \times_{t \in \{1,\ldots,T\}} \mathbb{X}_{tN}$. The subscript N allows for the regressors' support to change with the sample size. Available is a random sample $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^N$ from a distribution F_{0N}, which can depend on the sample size. The t^{th} period outcome for a random draw from F_{0N} is given by

$$Y_t = \mathbf{X}_t' B_t, \qquad t = 1, \ldots, T. \tag{2.1}$$

We assume that, conditional on $\mathbf{X}$, the P components of B_t are comonotonic:

$$(B_t | \mathbf{X} = \mathbf{x}) \overset{D}{=} \left(F^{-1}_{B_{1t}|\mathbf{X}}(V|\mathbf{x}), \ldots, F^{-1}_{B_{Pt}|\mathbf{X}}(V|\mathbf{x}) \right), \qquad V \sim \mathcal{U}[0,1]. \tag{2.2}$$

We also make the following common trends / stationarity assumption

$$F_{B_{p1}|\mathbf{X}}(b|\mathbf{x}) = F_{B_{pt}|\mathbf{X}}(b + \Delta_{pt}(b)|\mathbf{x}), \quad t = 2, \ldots, T, \quad p = 1, \ldots, P. \tag{2.3}$$

Solving for $\Delta_{pt}(b)$ yields

$$\Delta_{pt}(b) = F^{-1}_{B_{pt}|\mathbf{X}}\left(F_{B_{p1}|\mathbf{X}}(b|\mathbf{x}) \middle| \mathbf{x} \right) - b,$$

which after changing variables to $\tau = F_{B_{p1}|\mathbf{X}}(b|\mathbf{x})$ gives

$$\beta_{pt}(\tau; \mathbf{x}) - \beta_{p1}(\tau; \mathbf{x}) = \delta_{pt}(\tau),$$

for $\beta_{pt}(\tau; \mathbf{x}) = Q_{B_{pt}|\mathbf{X}}(\tau|\mathbf{x})$. Differences in the conditional quantile functions of B_{pt} and B_{ps} for $t \neq s$ do not depend on $\mathbf{X}$.

Let $Q_{Y_t|\mathbf{X}}(\tau|\mathbf{x})$ denote the τ^{th} conditional quantile function of Y_t given $\mathbf{X} = \mathbf{x}$; under (2.1), (2.2) and (2.3) we have

$$Q_{Y_t|\mathbf{X}}(\tau|\mathbf{x}) = \mathbf{x}_t' \left(\beta(\tau; \mathbf{x}) + \delta_t(\tau) \right), \quad t = 1, \ldots, T$$

where, $\beta(\tau; \mathbf{x}) = (\beta_{11}(\tau; \mathbf{x}), \ldots, \beta_{P1}(\tau; \mathbf{x}))'$, $\delta_t(\tau) = (\delta_{1t}(\tau), \ldots, \delta_{Pt}(\tau))'$ and we normalize

$\delta_1(\tau)$ to a vector of zeros. In an abuse of notation let

$$Q_{\mathbf{Y}|\mathbf{X}}(\tau|\mathbf{x}) = \begin{pmatrix} Q_{Y_1|\mathbf{X}}(\tau|\mathbf{x}) \\ Q_{Y_2|\mathbf{X}}(\tau|\mathbf{x}) \\ \vdots \\ Q_{Y_T|\mathbf{X}}(\tau|\mathbf{x}) \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \underline{0}_P' & \cdots & \underline{0}_P' \\ \mathbf{X}_2' & \cdots & \underline{0}_P' \\ \vdots & \ddots & \vdots \\ \underline{0}_P' & \cdots & \mathbf{X}_T' \end{pmatrix}, \quad \delta(\tau) = \begin{pmatrix} \delta_2(\tau) \\ \vdots \\ \delta_T(\tau) \end{pmatrix}.$$

Stacking equations we get

$$Q_{\mathbf{Y}|\mathbf{X}}(\tau|\mathbf{x}) = \mathbf{W}\delta(\tau) + \mathbf{X}\beta(\tau;\mathbf{X}). \tag{2.4}$$

More generally we study estimators based on (2.4) where $\mathbf{W}$ is a $T \times Q$ function of $\mathbf{X}$; in the benchmark model described above $Q = (T-1)P$.

2.2.1 Examples

Generalization of linear quantile regression model Let the t^{th} period outcome be given by

$$Y_t = \mathbf{X}_t'(\beta(U_t) + \delta_t), \qquad t = 1, \ldots, T \tag{2.5}$$

with $\mathbf{x}_t'\beta(u_t)$ increasing in the scalar u_t for all $u_t \in \mathbb{U}_t$ and all $\mathbf{x}_t \in \mathbb{X}_t$. We leave the conditional distribution of U_1 unrestricted, but assume marginal stationarity of U_t given $\mathbf{X}$ as in, for example, Manski (1987) :

$$U_1|\mathbf{X} \overset{D}{=} U_t|\mathbf{X}, \qquad t = 2, \ldots, T$$

Under marginal stationarity we may define the conditionally standard uniform random variable $V_t = F_{U_1|\mathbf{X}}(U_t|\mathbf{X})$ and rewrite (2.1):

$$\begin{aligned} Y_t &= \mathbf{X}_t'\left(\beta\left(F^{-1}_{U_1|\mathbf{X}}(V_t|\mathbf{X})\right) + \delta_t\right) \\ &= \mathbf{X}_t'(\beta(V_t;\mathbf{X}) + \delta_t), \qquad t = 1, \ldots, T. \end{aligned}$$

This gives

$$Q_{Y_t|\mathbf{X}}(\tau|\mathbf{x}) = \mathbf{x}_t'(\beta(\tau;\mathbf{x}) + \delta_t), \quad t = 1, \ldots, T.$$

This example is a very natural generalization of the random coefficient represention of the linear quantile regression model (e.g., Koenker (2005); pp. 59 - 62) to panel data. Note its fixed effects nature: the conditional distribution of U_1 given $\mathbf{X}$ is unrestricted. Here, the comonotoncity assumption on the random coefficients is equivalent to the assumption of one-dimensional heterogeneity.

Note that we can write $U_t = A + V_t$ and assume $V_1|\mathbf{X}, \mathbf{A} \overset{D}{=} V_t|\mathbf{X}, \mathbf{A}$ for $t = 2, \ldots, T$; this makes the 'fixed effects' nature of the model clearer.

Location-scale panel data model Let the t^{th} period outcome be given by

$$\begin{aligned} Y_t &= \mathbf{X}_t'\beta_t + \mathbf{X}_t'\gamma\left(U_t\right) \\ &= \mathbf{X}_t'\beta_t + \mathbf{X}_t'\gamma\left(A + V_t\right). \end{aligned} \tag{2.6}$$

with $\mathbf{x}_t'\gamma(u_t)$ increasing in u_t for all $u_t \in \mathbb{U}_t$ and all $\mathbf{x}_t \in \mathbb{X}_t$. This is equivalent to a comonotonicity assumption on the vector of coefficients $\gamma(\cdot)$.

Assume that the first element of $\mathbf{X}_t$ is a constant. Setting $\gamma = \left(1, \underline{0}_{P-1}'\right)$ yields the linear model studied by Chamberlain (1984). Maintaining marginal stationarity of U_t we get

$$\begin{aligned} Q_{Y_t|\mathbf{X}}\left(\tau|\,\mathbf{x}\right) &= \mathbf{x}_t'\left(\beta_t + \gamma Q_{A+V_1|\mathbf{X}}\left(\tau|\,\mathbf{x}\right)\right) \\ &= \mathbf{x}_t'\left(\beta\left(\tau;\mathbf{x}\right) + \delta_t\right), \end{aligned}$$

for $\beta\left(\tau;\mathbf{x}\right) = \beta_1 + \gamma Q_{A+V_1|\mathbf{X}}\left(\tau|\,\mathbf{x}\right)$ and $\delta_t = \beta_t - \beta_1$.

Both of the above models are more restrictive than the baseline set-up. This opens the door to potential testing of these extra restrictions.

2.2.2 Estimands

We focus on two main estimands in this chapter. The unconditional quantile effects (UQE) associated with the p^{th} regressor is

$$\beta_p\left(\tau\right) = \inf\left\{b_{p1} \in \mathbb{R} : F_{B_{p1}}\left(b_{p1}\right) \geq \tau\right\}. \tag{2.7}$$

If each individual in the population is given an additional unit of $\mathbf{X}_{p1}$, the effect on outcomes will be less than or equal to $\beta_p\left(\tau\right)$ for 100τ percent of the population. To get the corresponding object for a t^{th} period intervention we add $\delta_{pt}\left(\tau\right)$.

This is the quantile analog of an average partial effect (APE). Under our comonotonicity assumption this object is also related to differences in the quantile structural function (QSF) studied by Imbens and Newey (2009) and Chernozhukov et al. (2013).

We can generalize this idea to define 'decomposition effects' (cf., Melly (2006); Chernozhukov et al. (2009); Rothe (2010)) as well as deal with the case where the elements of $\mathbf{X}_t$ are functionally dependent (cf., Graham and Powell (2012)).

The second estimand is the average conditional quantile effect (ACQE). This object is similar to the average derivative quantile regression coefficients studied in Chaudhuri et al. (1997). It is also related to measures of average conditional inequality used in labor economics (e.g., Angrist et al. (2006); Lemieux (2006)). The $P \times 1$ vector of ACQEs in our model is given by

$$\overline{\beta}\left(\tau\right) = \mathbb{E}\left[\beta_1\left(\tau;\mathbf{X}\right)\right]. \tag{2.8}$$

2.3 Additional Assumptions

We now impose additional restrictions on the support of $\mathbf{X}$ and on the number of random coefficients.

2.3.1 Discrete Support

We first restrict our attention to the case where the support of $\mathbf{X}$ is discrete. The probabilities and support points are indexed by N, the sample size, and may depend on it.

When $\mathbf{X}_t$ is discrete, the support of the entire regressor sequence $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)'$ is finite: $\mathbf{X} \in \mathbb{X}_N = \{\mathbf{x}_{1N}, \ldots, \mathbf{x}_{MN}\}$. Without loss of generality assume that the first $m = 1, \ldots, L \leq M$ of these support points correspond to mover realizations of $\mathbf{X}$ such that $\operatorname{rank}(\mathbf{x}_{mN}) = P$. We further divide the mover realizations into "strict"-movers and "near"-stayers. A mover realization $\mathbf{x}_{mN}$ is a strict-mover if $\operatorname{rank}(\lim_{N\to\infty} \mathbf{x}_{mN}) = P$, and is a near-stayer if $\operatorname{rank}(\lim_{N\to\infty} \mathbf{x}_{mN}) < P$. Without loss of generality, let $m = 1, \ldots, L_1 \leq L$ denote the strict- movers and $m = L_1 + 1, \ldots, L$ denote the near-movers.

The remaining support points ($m = L+1, \ldots, M$) correspond to stayer realizations with $\operatorname{rank}(\mathbf{x}_{mN}) < P$. Let $\mathbb{X}_N^M = \{\mathbf{x}_{mN} : \operatorname{rank}(\mathbf{x}_{mN}) = P\}$ and $\mathbb{X}_N^S = \{\mathbf{x}_{mN} : \operatorname{rank}(\mathbf{x}_{mN}) < P\}$ be mutually exclusive subsets of $\mathbb{X}_N$ consisting of the mover and stayer support points respectively. Let $p_{mN} = \Pr(\mathbf{X} = \mathbf{x}_{mN})$ denote the probability associated with these support points.

In some cases, we will be unable to identify the UQE and ACQE as defined in the previous paragraphs. This can be due to the inability to identify certain characteristics of the distribution of the random ceofficients for stayers. Using this notation the movers UQE associated with the p^{th} element of X_1 is

$$\beta_{pN}^M(\tau) = \inf\left\{b_{p1} \in \mathbb{R} : F_{B_{p1}|\mathbf{X}\in\mathbb{X}_N^M}\left(b_{p1}|\,\mathbf{X} \in \mathbb{X}_N^M\right) \geq \tau\right\}, \tag{2.9}$$

while the corresponding movers ACQE is

$$\overline{\beta}_{pN}^M(\tau) = \mathbb{E}_N\left[\beta_{p1}(\tau; \mathbf{X})|\,\mathbf{X} \in \mathbb{X}_N^M\right]. \tag{2.10}$$

2.3.2 Just-identification and additional support assumptions

We also assume that the number of time periods, T is equal to the number of random coefficients P. With $T = P$, the matrix $\mathbf{X}$ is square, and therefore it has full rank if and only if $\det \mathbf{X} \neq 0$. We denote $\det \mathbf{X}$ as $D \in \mathbb{D}_N = \{\mathbf{d}_1, \ldots, \mathbf{d}_K, -h_N, h_N, 0\}$. We let $P_N(D = h_N) = P_N(D = -h_N) = bh_N$ for some $b \geq 0$, and we define $\mathbf{d}_{K+1} = -h_N$, $\mathbf{d}_{K+2} = h_N$ and $\mathbf{d}_{K+3} = 0$. We also let the probability of observing a singular $\mathbf{X}$ be $P_N(D = 0) = \pi_0^N = \pi_0 + 2bh_N$. Finally, $P_N(D = \mathbf{d}_k) = \pi_k^N$ for $k \in \{1, \ldots, K\}$ with $\sum_{k=1}^K \pi_K^N = 1 - 4bh_N - \pi_0$, with $4bh_N + \pi_0 < 1$ for all N. We also let $\pi_k = \lim_{N\to\infty} \pi_k^N$, so that $\sum_{k=0}^K \pi_k = 1$. Note that throughout the text, the absence of the N subscript on

probabilities or support points indicates that we are dealing with sequence limits, which we always assume exists.

In this setup, observations with $D = 0$ are stayers, $D = \pm h_N$ are near-stayers while $D = \mathbf{d}_k$ for $k = 1, \ldots, K$ denotes strict-mover realizations. The inclusion of near-stayers is a way to approximate a continuous distribution of D, letting some movers (those with $D = \pm h_N$) have very similar characteristics as stayers ($D = 0$).

We let $q_{mN|k} = P_N(\mathbf{X} = \mathbf{x}_{mN}|D = d_k)$, $q_{mN|-h} = P_N(\mathbf{X} = \mathbf{x}_{mN}|D = -h_N)$, $q_{mN|h} = P_N(\mathbf{X} = \mathbf{x}_{mN}|D = h_N)$ and $q_{mN|0} = P_N(\mathbf{X} = \mathbf{x}_{mN}|D = 0)$. For simplicity, we assume that $q_{mN|\cdot}$ does not vary with N, so that conditional on the value of the determinant, which has varying support, the distribution of $\mathbf{X}$ does not vary with the sample size. We also assume that $q_{m|h} = q_{m|-h} = q_{m|0}$ for all $m = 1, \ldots, M$.

We let either $\pi_0 > 0$ or $b > 0$ in this chapter. This means that there is always a non-zero fraction of observations that are stayers. If we allow $\pi_0 > 0$, that fraction is also asymptotically positive, which means that we can estimate some of their distributional features (i.e. time effects) with root-N consistent estimators. On the other hand, if we also have $b = 0$, we cannot learn about its some other features, (i.e. the random coefficients' distribution) since there are no movers nearby. In this setup, we will only be able to recover the movers' ACQE and UQE. We term this setup the "regular case".

On the other hand, if we have $b > 0$, $\pi_0 = 0$, the time effects will be estimated at rate root-Nh_N since the fraction of stayers is of order h_N, which implies a slower rate of convergence. The bandwidth case allows for using near-stayer observations to approximate stayers' behavior. This allows us to recover the true ACQE and UQE since we can approximate the stayers' contribution to the estimands with that of the near-stayers'. We use the "bandwidth case" ($b > 0$) as an approximation of the continuous case as in Graham and Powell (2012) since it allows for a singularity among stayer observations but also irregular (non root-N) identification of coefficients. This approximation simplifies calculations since the preliminary conditional quantile estimator for discrete conditioning variables is a straightforward one. In the presence of continuously-distributed covariates, there are estimators proposed in the literature (e.g. Belloni et al. (2011), Qu and Yoon (2011)) that have different desirable and undesirable features in the context of this chapter. The properties of the ACQE in the bandwidth case will have similar properties to the APE in Graham and Powell (2012), and here b will have a similar interpretation as ϕ_0 in Graham and Powell (2012), the density of the determinant evaluated at 0.

We do not consider here the case where both $\pi_0 > 0$ and $b > 0$, since in this case we need to estimate the stayers' random coefficient using a shrinking fraction of the sample size (near-stayers). This is similar to an estimation of a nonparametric derivative, and will suffer from a rate deterioration over the bandwidth case, since the ACQE and UQE would now converge at the rate root-Nh_N^3 instead of root-Nh_N.

Throughout we maintain the following assumptions:

Assumption 2.3.1 (Random Sampling) *$\{\mathbf{Y}_i, \mathbf{X}_i\}_{i=1}^N$ is a random sample from the population of interest. The distribution can vary with the sample size.*

Assumption 2.3.2 (Bounded and Continuous Densities) *The conditional distribution* $F_{Y_t|X}(y_t|\mathbf{x})$ *has density* $f_{Y_t|X}(y_t|\mathbf{x})$ *such that* $\phi(\tau;\mathbf{x}) = f_{Y_t|X}\left(F^{-1}_{Y_t|X}(\tau|\mathbf{x})\middle|\mathbf{x}\right)$ *and* $\phi'(\tau;\mathbf{x})$ *are uniformly bounded for all* $\tau \in (0,1)$*, all* $\mathbf{x} \in \mathbb{X}_N$*, and all* $t = 1,\ldots,T$*. Also, this conditional distribution does not vary with the sample size* N*. Finally,* $f_{Y_t|X}(y_t|\mathbf{x})$*,* $F_{Y_t|X}(y_t|\mathbf{x})$ *and* $F^{-1}_{Y_t|X}(y_t|\mathbf{x})$ *are all continuous in* $\mathbf{x}$*.*

Assumption 2.3.3 (Bounded $\mathbf{Y}$) *The support of* $\mathbf{Y}_t$ *is compact for all* $t \in \{1,\ldots,T\}$*.*

Let

$$\widehat{F}_{Y_t|\mathbf{X}}(y_t|\mathbf{x}_{mN}) = \left[\sum_{i=1}^{N}\mathbf{1}(\mathbf{X}_i = \mathbf{x}_{mN})\right]^{-1} \times \left[\sum_{i=1}^{N}\mathbf{1}(\mathbf{X}_i = \mathbf{x}_{mN})\mathbf{1}(Y_{it} \leq y_t)\right],$$

be the empirical cumulative distribution function of Y_t for the subsample of units with $\mathbf{X} = \mathbf{x}_{mN}$. We estimate the τ^{th} conditional quantile of Y_t by

$$\widehat{\mathbf{\Pi}}_t(\tau;\mathbf{x}_{mN}) = \widehat{F}^{-1}_{Y_t|\mathbf{X}}(y_t|\mathbf{x}_{mN}) = \inf\left\{y_t : \widehat{F}_{Y_t|\mathbf{X}}(y_t|\mathbf{x}_m) \geq \tau\right\},$$

where $\mathbf{\Pi}_t(\tau;\mathbf{X})$ denotes the t^{th} element of $\mathbf{\Pi}(\tau;\mathbf{X})$. Note that $\widehat{\mathbf{\Pi}}(\cdot;\mathbf{x}_{lN})$ and $\widehat{\mathbf{\Pi}}(\cdot;\mathbf{x}_{mN})$ for $l \neq m$ are conditionally uncorrelated given $\{\mathbf{X}\}_{i=1}^{N}$.

Define

$$\rho_{st}(\tau,\tau';\mathbf{X}) = \frac{\Pr(Y_{is} \leq \mathbf{\Pi}_s(\tau;\mathbf{X}), Y_{it} \leq \mathbf{\Pi}_t(\tau;\mathbf{X})) - \tau\tau'}{\min(\tau,\tau') - \tau\tau'}, \quad s,t = 1,\ldots,T \tag{2.11}$$

and

$$\Lambda(\tau,\tau';\mathbf{X}) = \begin{pmatrix} \frac{1}{f_{Y_1|\mathbf{X}}(\mathbf{\Pi}_1(\tau;\mathbf{X}))f_{Y_1|\mathbf{X}}(\mathbf{\Pi}_1(\tau';\mathbf{X}))} & \cdots & \frac{\rho_{1T}(\tau,\tau';\mathbf{X})}{f_{Y_1|\mathbf{X}}(\mathbf{\Pi}_1(\tau;\mathbf{X}))f_{Y_T|\mathbf{X}}(\mathbf{\Pi}_T(\tau';\mathbf{X}))} \\ \vdots & \ddots & \vdots \\ \frac{\rho_{1T}(\tau,\tau';\mathbf{X})}{f_{Y_T|\mathbf{X}}(\mathbf{\Pi}_T(\tau;\mathbf{X}))f_{Y_1|\mathbf{X}}(\mathbf{\Pi}_1(\tau';\mathbf{X}))} & \cdots & \frac{1}{f_{Y_T|\mathbf{X}}(\mathbf{\Pi}_T(\tau;\mathbf{X}))f_{Y_T|\mathbf{X}}(\mathbf{\Pi}_T(\tau';\mathbf{X}))} \end{pmatrix}. \tag{2.12}$$

Using this notation, an adaptation of standard results on quantile processes in the cross sectional context, gives our first result:

Proposition 2.3.4 *Suppose that Assumptions 2.3.1, 2.3.2 and 2.3.3 are satisfied, then* $\sqrt{Np_{mN}}\left(\widehat{\mathbf{\Pi}}(\tau;\mathbf{x}_{mN}) - \mathbf{\Pi}(\tau;\mathbf{x}_{mN})\right)$ *converges in distribution to a mean zero Gaussian process* $\mathbf{Z}_Q(\cdot,\cdot)$ *on* $\tau \in (0,1)$ *and* $\mathbf{x}_{mN} \in \mathbb{X}_{\mathbb{N}}$*, where* $\mathbf{Z}_Q(\cdot,\cdot)$ *is defined by its covariance function* $\Sigma(\tau,\mathbf{x}_l,\tau',\mathbf{x}_m) = \mathbb{E}\left[\mathbf{Z}_Q(\tau,\mathbf{x}_l)\mathbf{Z}_Q(\tau',\mathbf{x}_m)'\right]$ *with*

$$\Sigma(\tau,\mathbf{x}_l,\tau',\mathbf{x}_m) = (\min(\tau,\tau') - \tau\tau')\Lambda(\tau,\tau';\mathbf{x}_l)\cdot\mathbf{1}(l=m)$$

for $l,m = 1,\ldots,L$*.*

These results are a standard generalization of process convergence results for unconditional quantiles. Convergence for support points with probability that shrink to 0 at rate h_N will be of order root-Nh_N since the is effective sample size used to estimate this conditional quantile is proportional to Nh_N rater than N. Convergence here is on $\tau \in (0,1)$ since we assume that Y_t has compact support. If Y_t's support was unbounded, results would instead hold uniformly on $\tau \in [\epsilon, 1-\epsilon]$ for arbitrary ϵ satisfying $0 < \epsilon < 1/2$. We now proceed with the identification and estimation of the time effect $\delta(\cdot)$. Let $\mathbf{X}^*$ denote the adjoint matrix of $\mathbf{X}$. From equation (2.4), we have that

$$\mathbf{X}^* Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}) = \mathbf{W}^* \delta(\tau) + D\beta(\tau; \mathbf{X}),$$

where $\mathbf{W}^* = \mathbf{X}^*\mathbf{W}$. Let $\mathbf{x}_{lN}$ be a stayer realization ($\mathbf{x}_{lN} \in \mathbb{X}_N^S$), therefore $l \in \{L+1, \ldots, M\}$ and $D = 0$. Then, for any $l \in \{L+1, \ldots, M\}$, we have:

$$\mathbf{w}^{*\prime}_{lN} \mathbf{x}^*_{lN} Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_{lN}) = \mathbf{w}^{*\prime}_{lN} \mathbf{w}^*{}_{lN} \delta(\tau). \tag{2.13}$$

The time effect is over identified if $L+1 < M$, and we can recover it in multiple ways. We choose here to average over the stayer realizations:

$$\delta(\tau) = \left(\sum_{l=L+1}^{M} \mathbf{w}^{*\prime}_{lN} \mathbf{w}^*{}_{lN} p_{lN} \right)^{-1} \sum_{l=L+1}^{M} \mathbf{w}^{*\prime}_{lN} \mathbf{x}^*_{lN} Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_{lN})\, p_{lN}. \tag{2.14}$$

Note that this expression is also equal to

$$\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^* | D = 0\right]^{-1} \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^* Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X})\,|D = 0\right].$$

We then define the analog estimator:

$$\widehat{\delta}(\tau) = \left(\sum_{l=L+1}^{M} \mathbf{w}^{*\prime}_{lN} \mathbf{w}^*{}_{lN} \hat{p}_{lN} \right)^{-1} \sum_{l=L+1}^{M} \mathbf{w}^{*\prime}_{lN} \mathbf{x}^*_{lN} \widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_{lN})\, \hat{p}_{lN}. \tag{2.15}$$

where $\hat{p}_{lN} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\mathbf{X}_i = \mathbf{x}_{lN})$ is the empirical probability associated with the lth realization. This estimator can also be written down as:

$$\widehat{\delta}(\tau) = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{W}^{*\prime}_i \mathbf{W}^*{}_i \mathbf{1}(|D_i| < h_N) \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{W}^{*\prime}_i \mathbf{X}^*_i \widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\, \mathbf{1}(D_i < h_N). \tag{2.16}$$

Proposition 2.3.5 *Let $Nh_N \to \infty$ as $N \to \infty$. Under assumptions 2.3.1, 2.3.2 and 2.3.3, we have that*

$$\sqrt{N\pi_0^N}\left(\widehat{\delta}\left(\tau\right)-\delta\left(\tau\right)\right)\xrightarrow{d}\mathbf{Z}_\delta(\tau)$$

on $\tau\in(0,1)$*, with*

$$\begin{aligned}\mathbb{E}\left[\mathbf{Z}_\delta(\tau)\mathbf{Z}_\delta(\tau')'\right]=(\min(\tau,\tau')-\tau\tau')\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^{*}|D=0\right]^{-1}\times\\ \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^{*}\Lambda(\tau,\tau';\mathbf{X})\mathbf{X}^{*\prime}\mathbf{W}^{*}|D=0\right]\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^{*}|D=0\right]^{-1}.\end{aligned}$$

Note that this result holds no matter if $\pi_0>0$ or $\pi_0=0$. In the case where $\pi_0>0$, this estimator is root-N consistent since we are using a non-shrinking fraction of observations in our estimate. When $b>0$ and π_0, we are using a number of observations approximately equal to $2bNh_N$ which makes the rate of convergence equal to root-Nh_N.

Having identified the time effect $\delta(\tau)$, we can now recover the conditional distribution of the coefficient of interest. Again premultiplying the model by the adjoint matrix of $\mathbf{X}$, we get:

$$\mathbf{X}^{*}Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}\right)=\mathbf{W}^{*}\delta\left(\tau\right)+D\beta\left(\tau;\mathbf{X}\right).$$

The conditional β can be recovered for both strict and near movers ($D\neq 0$) in the following way:

$$\beta\left(\tau;\mathbf{X}\right)=\frac{\mathbf{X}^{*}Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}\right)-\mathbf{W}^{*}\delta\left(\tau\right)}{D}$$

We now turn to the estimation of two different functionals of the conditional beta: the ACQE and the UQE.

2.4 Estimation of the ACQE

We again distinguish between the regular case where $\pi_0>0$ and $b=0$, and the bandwidth case where $\pi_0=0$ and $b>0$, as these will have different rates of convergence and asymptotic variances.

2.4.1 ACQE in the regular case

In this case, we have a fixed fraction of stayers and no near-stayers. We focus on the estimation of the movers' ACQE, defined in (2.10). The estimator we propose averages estimates for conditional betas using sample probabilities. The conditional betas estimates

also rely on estimates for $\delta(\cdot)$, the time effect. We first compute the asymptotic distribution of an infeasible estimator, computed assuming $\delta(\cdot)$ is known. We assume for simplicity that all support points and probabilities do not vary with N, the sample size.

Proposition 2.4.1 *Let*

$$\begin{aligned}\widehat{\widehat{\beta}}_I^M(\tau) &= \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i) - \mathbf{W}_i\delta(\tau)\right)\mathbf{1}(\mathbf{X}_i \in \mathbb{X}^M)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in \mathbb{X}^M)} \\ &= \sum_{l=1}^{L_1} \mathbf{x}_l^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l) - \mathbf{W}_l\delta(\tau)\right)\widehat{q}_l^M\end{aligned}$$

be the infeasible movers' ACQE, where $q_l^M = \frac{p_l}{1-\pi_0}$*, is the conditional probability of realization* l *conditional on being a mover realization, and let* $\widehat{q}_l^M = \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i=\mathbf{x}_l)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\det(\mathbf{X}_i)\neq 0)}$*. Under assumptions 2.3.1, 2.3.2 and 2.3.3, we have that:*

$$\sqrt{N}\left(\widehat{\widehat{\beta}}_I^M(\tau) - \bar{\beta}^M(\tau)\right) \xrightarrow{d} Z_I(\tau)$$

on $\tau \in (0,1)$ *for* $Z_I(\cdot)$ *a gaussian process with covariance equal to*

$$\begin{aligned}\mathbb{E}\left[Z_I(\tau)Z_I(\tau')'\right] &= \Psi_1(\tau,\tau') + \Psi_2(\tau,\tau') \\ \Psi_1(\tau,\tau') &= \frac{1}{1-\pi_0} Cov\left(\beta(\tau,\mathbf{X}), \beta(\tau',\mathbf{X})|X \in \mathbb{X}^M\right) \\ \Psi_2(\tau,\tau') &= \frac{\min(\tau,\tau') - \tau\tau'}{1-\pi_0}\mathbb{E}\left[\mathbf{X}^{-1}\Lambda(\tau,\tau',\mathbf{X})\mathbf{X}^{-1\prime}|\mathbf{X} \in \mathbb{X}^M\right].\end{aligned}$$

The form of the covariance function in Theorem 2.4.1 mirrors the basic form found by Chamberlain (1992) for the average partial effect estimand (see also Graham and Powell (2012)). This parallel is easiest to see for the case where $\tau = \tau'$. In that case the first component $\Psi_1(\tau,\tau)$ equals the variance (over $\mathbf{X}$) of the conditional quantile effects $\beta(\tau;\mathbf{X})$. The second term $\Psi_2(\tau,\tau)$ equals the average of the variances of the individual CQE estimates $\widehat{\beta}(\tau;\mathbf{X})$ when $\delta(\tau)$ is known. The final term, not included here, captures the asymptotic penalty associated with having to estimate $\delta(\tau)$. Since this is the infeasible ACQE, this penalty does not affect the estimate. It is also of note that this estimator is independent of $\widehat{\delta}(\cdot)$, since we are using non-overlapping subsamples (units with $D \neq 0$ and $D = 0$,

respectively) to compute the respective estimates. We now turn our attention to the feasible ACQE and compute its asymptotic variance.

Proposition 2.4.2 *Let*

$$\begin{aligned}\widehat{\bar{\beta}}^{M}(\tau) &= \frac{\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)-\mathbf{W}_i\widehat{\delta}(\tau)\right)\mathbf{1}(\mathbf{X}_i\in\mathbb{X}^M)}{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\mathbf{X}_i\in\mathbb{X}^M)}\\ &= \sum_{l=1}^{L_1}\mathbf{x}_l^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l)-\mathbf{W}_l\widehat{\delta}(\tau)\right)\widehat{q}_l^{M}\end{aligned}$$

be the feasible movers' ACQE, where $q_l^M=\frac{p_l}{1-\pi_0}$*, is the conditional probability of realization* l *conditional on being a mover realization, and let* $\widehat{q}_l^M=\frac{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\mathbf{X}_i=\mathbf{x}_l)}{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\det(\mathbf{X}_i)\neq 0)}$*. Under assumptions 2.3.1, 2.3.2 and 2.3.3, we have that:*

$$\sqrt{N}\left(\widehat{\bar{\beta}}^{M}(\tau)-\bar{\beta}^{M}(\tau)\right)\xrightarrow{d}Z(\tau)=Z_I(\tau)+\frac{\Xi_0}{\sqrt{\pi_0}}Z_\delta(\tau)$$

on $\tau\in(0,1)$*, and* $\Xi_0=\mathbb{E}\left[\mathbf{X}^{-1}\mathbf{W}|\mathbf{X}\in\mathbb{X}^M\right]$*. The variance of the gaussian process* $Z(\cdot)$ *is defined as*

$$\begin{aligned}\mathbb{E}\left[Z(\tau)Z(\tau')'\right] &= \Psi_1(\tau,\tau')+\Psi_2(\tau,\tau')+\Xi_0\Psi_3(\tau,\tau')\Xi_0'\\ \Psi_1(\tau,\tau') &= \frac{1}{1-\pi_0}Cov\left(\beta(\tau,\mathbf{X}),\beta(\tau',\mathbf{X})|X\in\mathbb{X}^M\right)\\ \Psi_2(\tau,\tau') &= \frac{\min(\tau,\tau')-\tau\tau'}{1-\pi_0}\mathbb{E}\left[\mathbf{X}^{-1}\Lambda(\tau,\tau',\mathbf{X})\mathbf{X}^{-1\prime}|\mathbf{X}\in\mathbb{X}^M\right]\\ \Psi_3(\tau,\tau') &= \frac{\min(\tau,\tau')-\tau\tau'}{\pi_0}\times\\ &\qquad\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^{*}|D=0\right]^{-1}\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^{*}\Lambda(\tau,\tau';\mathbf{X})\mathbf{X}^{*\prime}\mathbf{W}^{*}|D=0\right]\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^{*}|D=0\right]^{-1}.\end{aligned}$$

The asymptotic variance of the feasible estimator differs from the infeasible term due to the presence of term $\Psi_3(\cdot)$, which is the asymptotic variance of $\sqrt{N}(\widehat{\delta}(\cdot)-\delta(\cdot))$, weighted by Ξ_0.

We have established the identification and a consistent estimate for the movers' ACQE in the case where there is a point mass of stayers, but no near-stayers to help identify the stayer's main coefficient $\beta(\tau;\mathbf{X})$. We now focus on the identification and estimation of the

true ACQE in the bandwidth case, where there is a shrinking fraction of stayers and a equal amount of near-stayers.

2.4.2 ACQE in the bandwidth case

First note, that when $\pi_0 = 0$, the time effect estimator $\widehat{\delta}(\tau)$ is root-Nh_N consistent rather than root-N due to the shrinking fraction of the sample used to estimate it. We use the same estimator for the ACQE that is used in the case where $\pi_0 > 0$ and $b = 0$. In this case the difference between the ACQE and the movers' ACQE will be $O(h_N)$ meaning that it will asymptotically disappear under appropriate rate assumptions on h_N. Here is the main result:

Proposition 2.4.3 *Let*

$$\begin{aligned}\widehat{\bar{\beta}}_N(\tau) &= \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right) - \mathbf{W}_i\widehat{\delta}(\tau)\right)\mathbf{1}(|D_i| \geq h_N)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| \geq h_N)} \\ &= \sum_{l=1}^L \mathbf{x}_l^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_l\right) - \mathbf{W}_l\widehat{\delta}(\tau)\right)\widehat{q}_l^M\end{aligned}$$

be the feasible movers' ACQE, where $q_l^M = \frac{p_l}{1-\pi_0}$, is the conditional probability of realization l conditional on being a mover realization, and let $\widehat{q}_l^M = \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i = \mathbf{x}_l)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| \geq h_N)}$. Under assumptions 2.3.1, 2.3.2 and 2.3.3 , and if $Nh_N \to \infty$ and $Nh_N^3 \to 0$ as $N \to \infty$, we have that:

$$\sqrt{Nh_N}\left(\widehat{\bar{\beta}}(\tau) - \bar{\beta}_N(\tau)\right) \xrightarrow{d} Z(\tau)$$

for $\tau \in (0,1)$, and

$$\begin{aligned}\mathbb{E}\left[Z(\tau)Z(\tau')'\right] &= \Psi_1(\tau,\tau') + \Xi_0\Psi_2(\tau,\tau')\Xi_0' \\ \Psi_1(\tau,\tau') &= 2b\mathbb{E}\left[\mathbf{X}^*\Lambda(\tau,\tau',\mathbf{X})\mathbf{X}^{*\prime}|D=0\right] \\ \Psi_2(\tau,\tau') &= \frac{1}{2b}\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^*|D=0\right]^{-1} \times \\ &\qquad \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^*\Lambda(\tau,\tau',\mathbf{X},\mathbf{X})\mathbf{X}^{*\prime}\mathbf{W}^*|D=0\right]\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^*|D=0\right]^{-1},\end{aligned}$$

where $\Xi_0 = \lim_{N\to\infty}\mathbb{E}_N\left[\mathbf{X}^{-1}\mathbf{W}|\mathbf{X} \in \mathbb{X}_N^M\right]$.

We can interpret the asymptotic results in the following way: there is a bias term of order $O(h_N)$ due to trimming of stayers, which represent a fraction proportional to h_N of the sample. This vanishes since Nh_N^3 goes to 0 as N goes to infinity.

There is also some randomness due to the random coefficient itself, $\beta(\tau; \mathbf{X})$. This term vanishes asymptotically, since it is of order $O_p\left(\frac{1}{\sqrt{N}}\right)$. This is in contrast to the framework with non-shrinking probabilities on all probability points, where the order of convergence is root-N everywhere. We could potentially improve asymptotic properties of a variance estimate for $\widehat{\bar{\beta}}(\cdot)$ by correcting for this lower order term. In this framework, this source of noise is dominated by the variance coming from the estimation of the conditional betas for near-stayers.

This variance, coming from $\Psi_1(\cdot, \cdot)$ is due to estimation error of the conditional quantiles. It can be further decomposed in the estimation error of conditional quantiles for strict-movers, and that for near-stayers. Since there is a non-shrinking fraction of strict-movers, the conditional quantiles for strict-movers can be estimated at root-N rate and, when these quantiles are averaged using the sampling distribution of $\mathbf{X}$, their contribution remain of the order root-N.

The contribution of near-stayers to the variance is analogous to the reason the estimator $\widehat{\beta}^I$ in Graham and Powell (2012) converges at the rate root-Nh_N. The conditional quantiles for near-stayers is estimated at rate root-Nh_N since there is a shrinking fraction (equal to $2bh_N$) of near-stayers. These quantiles are premultiplied by $\mathbf{X}^{-1} = \frac{\mathbf{X}^*}{D}$, and since $D = \pm h_N$, this makes the quantile error divided by the determinant of order $O_p\left(\frac{1}{\sqrt{Nh_N^3}}\right)$, which is a problem since we need to assume $Nh_N^3 \to 0$ to get rid of the bias term. Fortunately, this problem goes away since the quantile error divided by the determinant is weighted by the empirical probability for these near-movers, which is of order $O(h_N)$. Combining these terms, the contribution of near-movers to the infeasible ACQE remains root-Nh_N. This is the dominant term in the asymptotic expansion of this estimator when using the bandwidth framework.

The term $\Psi_2(\cdot, \cdot)$ is the contribution of $\widehat{\delta}(\cdot)$ to the estimation error. The errors from the estimation of conditional quantiles for movers and from the estimation of time effects are independent since they rely on different subsamples, making their linear combination's asymptotic variance easier to compute. The estimation error of the time effect is premultiplied by $\widehat{\Xi}_N = \frac{1}{N}\sum_{i=1}^N \frac{\mathbf{W}_i^* \mathbf{1}(|D_i| \geq h_N)}{D_i}$, which converges in probability to Ξ_0, a well defined limit under the assumptions here. This is analogous to the structure of the APE in Graham and Powell (2012).

2.5 Estimation of the UQE

In this section, we derive a consistent and asymptotically normal estimator for the UQE (as defined in 2.7) and for the movers' UQE (defined in 2.9) processes, in the bandwidth and regular discrete case, respectively. We first focus on the regular discrete case and the

estimation of the movers' UQE.

2.5.1 UQE in the Regular Case

In this case, let $\widehat{\beta}_p^M(\tau)$ be an estimate of the pth component of the movers' UQE at quantile τ defined by

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\mathbf{X}_i \in \mathbb{X}^M\right)\int_0^1\left(\mathbf{1}\left(\widehat{\beta}_p(u,\mathbf{X}_i) \leq \widehat{\beta}_p^M(\tau)\right)-\tau\right)du = 0. \tag{2.17}$$

We assume here that the integral over u can be computed without approximation required. This can be justified by the use of interpolation around a finite number of points to compute the conditional beta's process' estimate, which means that an integral could potentially be computed exactly. This estimator is the τth quantile of the empirical distribution of $\beta_p(U,\mathbf{X})$ given that $\mathbf{X} \in \mathbb{X}^M$, which is approximated by the distribution of $\widehat{\beta}_p(U,\mathbf{X})$. To derive the estimator's asymptotic properties, we first compute the asymptotic distribution of the CDF of $\widehat{\beta}_p(U,\mathbf{X})$ and then invert it at $\beta_p^M(\tau)$. The actual estimate for the movers' UQE is computed in a similar fashion using a CDF estimate.

Proposition 2.5.1 *Let $p \in \{1,\dots,P\}$. Then we have that*

$$\sqrt{N}\left(\widehat{\beta}_p^M(\tau)-\beta_p^M(\tau)\right) \xrightarrow{d} Z_{UQE}(\tau)$$

on $\tau \in (0,1)$ with $Z_{UQE}(\cdot)$ being a Gaussian process. Let $d = F_{\mathbf{B}_p|\mathbf{X}}(\beta_p^M(\tau),\mathbf{X})$ and $d' = F_{\mathbf{B}_p|\mathbf{X}}(\beta_p^M(\tau'),\mathbf{X})$. The covariance of this Gaussian process is equal to:

$$\begin{aligned}
\mathbb{E}\left[Z_{UQE}(\tau)Z_{UQE}(\tau')'\right] &= \frac{\Psi_1(\tau,\tau')+\Psi_2(\tau,\tau')+\Psi_3(\tau,\tau')}{f_{\mathbf{B}_p|\mathbf{X}\in\mathbb{X}^M}(\beta_p^M(\tau))f_{\mathbf{B}_p|\mathbf{X}\in\mathbb{X}^M}(\beta_p^M(\tau'))}\\
\Psi_1(\tau,\tau') &= \mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(\beta_p^M(\tau),\mathbf{X})f_{\mathbf{B}_p|\mathbf{X}}(\beta_p^M(\tau'),\mathbf{X})\mathbf{X}^{-1}(\min(d,d')-dd')\times\right.\\
&\quad \left.\Lambda(d,d',\mathbf{X})\mathbf{X}^{-1\prime}|\mathbf{X}\in\mathbb{X}^M\right]\\
\Psi_2(\tau,\tau') &= \frac{1}{\pi_0}\mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(d,\mathbf{X})\mathbf{X}^{-1}\mathbf{W}Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(\beta_p^M(\tau),\mathbf{X}))\times\right.\\
&\quad \left.Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(\beta_p^M(\tau'),\tilde{\mathbf{X}}))'\tilde{\mathbf{W}}'\tilde{\mathbf{X}}^{-1\prime}f_{\mathbf{B}_p|\mathbf{X}}(d',\tilde{\mathbf{X}})|\mathbf{X}\in\mathbb{X}^M,\tilde{\mathbf{X}}\in\mathbb{X}^M\right]\\
\Psi_3(\tau,\tau') &= \frac{1}{1-\pi_0}Cov\left(d,d'|\mathbf{X}\in\mathbb{X}^M\right)
\end{aligned}$$

where $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are independent copies.

We can see that the asymptotic variance of the UQE also has a three term decomposition with a similar interpretation to the ACQE. The first term $\Psi_1(\cdot,\cdot)$ represents the uncertainty due to to the estimation of the conditional quantiles for movers, while the second term $\Psi_2(\cdot,\cdot)$ is due to the estimation error of the time effect. Together, they give the estimation error of estimating conditional betas for movers. Finally, $\Psi_3(\cdot,\cdot)$ represents the inherent error due to the fact that coefficients are random and correlated with $\mathbf{X}$. Again, we can see that these three terms are due to three independent sources of variation: the variation in $\mathbf{Y}$ conditional on $\mathbf{X}$ being a mover realization, the variation in $\mathbf{Y}$ conditional on $\mathbf{X}$ being a stayer realization, and the variation in $\mathbf{X}$ itself. These three sources lead to $\Psi_1(\cdot,\cdot)$, $\Psi_2(\cdot,\cdot)$ and $\Psi_3(\cdot,\cdot)$ respectively. We now consider the asymptotic distribution of the UQE in the bandwidth case:

2.5.2 UQE in the Bandwidth Case

We define the same estimator as in the regular case, which inverts the unconditional beta's CDF at quantile τ. For this estimator, asymptotic rate of convergence will be root-Nh_N as in the ACQE's case, and its asymptotic variance also contains two terms. Here is the main proposition.

Proposition 2.5.2 *Let $p \in \{1,\dots,P\}$, $Nh_N \to \infty$ and $Nh_N^3 \to 0$ as $N \to \infty$. Then we have that*

$$\sqrt{Nh_N}\left(\widehat{\beta}_{pN}(\tau) - \beta_{pN}(\tau)\right) \xrightarrow{d} Z_{UQE}(\tau)$$

on $\tau \in (0,1)$ with $Z_{UQE}(\cdot)$ being a Gaussian process. Let $d = F_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau),\mathbf{X})$ and $d' = F_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau'),\mathbf{X})$. The covariance of this Gaussian process is equal to:

$$\begin{aligned}
\mathbb{E}\left[Z_{UQE}(\tau)Z_{UQE}(\tau')'\right] &= \frac{\Psi_1(\tau,\tau') + \Psi_2(\tau,\tau')}{f_{\mathbf{B}_p}(\beta_p(\tau))f_{\mathbf{B}_p}(\beta_p(\tau'))} \\
\Psi_1(\tau,\tau') &= 2b\mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau),\mathbf{X})f_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau'),\mathbf{X})\mathbf{X}^*\times\right. \\
&\qquad \left.(\min(d,d') - dd')\Lambda(d,d',\mathbf{X})\mathbf{X}^{*\prime}|D=0\right] \\
\Psi_2(\tau,\tau') &= 2b\mathbb{E}\Big[f_{\mathbf{B}_p|\mathbf{X}}(d,\mathbf{X})\mathbf{W}^*Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau),\mathbf{X}))Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau'),\tilde{\mathbf{X}}))'\times \\
&\qquad \tilde{\mathbf{W}}^{*\prime}f_{\mathbf{B}_p|\mathbf{X}}(d',\tilde{\mathbf{X}})|D=0,\tilde{D}=0\Big]
\end{aligned}$$

where $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are independent copies.

The asymptotic variance of the estimator contains two terms, the first relating to the estimation of conditional quantiles for near-stayers, while the second term reflects the esti-

mation error related to the time effect. The strict movers have no influence on the asymptotic variance since their contribution is of order $O_p\left(\frac{1}{\sqrt{N}}\right)$. It would be again possible to get small sample improvements in the variance estimate by using a correction for this lower order term. The term reflecting the variance of the random coefficient is washed away here, since it is also of order $O_p\left(\frac{1}{\sqrt{N}}\right)$. This result is analogous to the ACQE's variance in the bandwidth case. There is also a bias term of order $O(h_N)$ that vanishes as $Nh_N^3 \to 0$. This bias arises from estimating the movers' UQE, which converges to the true UQE at rate $O(h_N)$.

2.6 Conclusion

We have derived the asymptotic distribution of the ACQE and the UQE in a just-identified panel data model with correlated random coefficients. In the regular case, we show the root-N consistency for estimators of the ACQE and UQE, and compute the asymptotic variance of these estimators. We also consider the bandwidth case as an approximation for the continuous case, then show the root-Nh_N consistency for estimators of the ACQE and UQE. We then compute the asymptotic variance of these estimators. The small-sample properties of our estimators and the distribution of the estimators when covariates are continuously distributed represent areas of future research.

2.7 Proofs of Propositions

Proof of proposition 2.3.4. See Proposition 2.1 in Graham-Hahn-Powell (2011). We instead use Lyapunov's central limit theorem for support points and probabilities that are indexed by N. ■

Lemma 2.7.1 *In the bandwidth discrete case,*

$$\frac{1}{Nh_N}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_i\mathbf{W}^{*}{}_i\mathbf{1}(|D_i| < h_N) \xrightarrow{\mathbb{P}} 2\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^{*}|D=0\right]b$$

where $P_N(D=0) = 2bh_N$.

Proof.

$$\begin{aligned}\frac{1}{Nh_N}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_i\mathbf{W}^{*}{}_i\mathbf{1}(|D_i|<h_N) &= 2b\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{lN}\mathbf{w}^{*}{}_{lN}\frac{\hat{p}_{lN}}{P_N(D=0)}\\ &\xrightarrow{\mathbb{P}} 2b\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{w}^{*}{}_{l}q_{l|0}\\ &= 2\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^{*}|D=0\right]b\end{aligned}$$

■

Lemma 2.7.2 *In the bandwidth discrete case,*

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_i\mathbf{X}_i^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)\right)\mathbf{1}(|D_i|<h_N)\xrightarrow{d}Z_1(\tau)$$

with $\mathbb{E}\left[Z_1(\tau)Z_1(\tau')'\right]=2b\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^{*}\Sigma(\tau,\tau',\mathbf{X},\mathbf{X})\mathbf{X}^{*\prime}\mathbf{W}^{*}|D=0\right]$.

Proof.

$$\begin{aligned}&\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_i\mathbf{X}_i^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)\right)\mathbf{1}(|D_i|<h_N)\\ &\qquad=\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_i\mathbf{X}_i^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)\right)\mathbf{1}(D_i=0)\\ &\qquad=\sqrt{\frac{N}{h_N}}\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{lN}\mathbf{x}_{lN}^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_{lN}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_{lN}\right)\right)\hat{p}_{lN}\\ &\qquad=\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{lN}\mathbf{x}_{lN}^{*}\sqrt{Np_{lN}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_{lN}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_{lN}\right)\right)\frac{\hat{p}_{lN}}{\sqrt{p_{lN}h_N}}\\ &\qquad=\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{lN}\mathbf{x}_{lN}^{*}\sqrt{Np_{lN}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_{lN}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{x}_{lN}\right)\right)\times\\ &\frac{\hat{p}_{lN}}{\sqrt{p_{lN}P_N(D=0)}}\sqrt{2b}\\ &\xrightarrow{d}\sqrt{2b}\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{x}_{l}^{*}Z_Q(\tau,\mathbf{x}_l)\sqrt{q_{l|0}}\\ &=Z_1(\tau)\end{aligned}$$

Since $E[Z_Q(\tau, \mathbf{x}_l) Z_Q(\tau', \mathbf{x}_m)'] = \Sigma(\tau, \tau', \mathbf{x}_l, \mathbf{x}_m) = (\min(\tau, \tau') - \tau\tau')\Lambda(\tau, \tau'; \mathbf{x}_l) \cdot \mathbf{1}(l = m)$, we get that:

$$\begin{aligned}\mathbb{E}[Z_1(\tau) Z_1(\tau')] &= 2b \sum_{l=L+1}^{M} \mathbf{W}_{\mathbf{l}}^{*\prime} \mathbf{x}_l^* \Sigma(\tau, \tau', \mathbf{x}_l, \mathbf{x}_l) \mathbf{x}_l^{*\prime} \mathbf{W}_{\mathbf{l}}^* q_{l|0} \\ &= 2b\mathbb{E}\left[\mathbf{W}^{*\prime} \mathbf{X}^* \Sigma(\tau, \tau', \mathbf{X}, \mathbf{X}) \mathbf{X}^{*\prime} \mathbf{W}^* | D = 0\right].\end{aligned}$$

■

Lemma 2.7.3 *In the regular discrete case,*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{W}^{*\prime}{}_i \mathbf{W}^*{}_i \mathbf{1}(D_i = 0) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\mathbf{W}^{*\prime} \mathbf{W}^* | D = 0\right] \pi_0.$$

Proof. Straightforward use of LLN. ■

Lemma 2.7.4 *In the regular discrete case,*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{W}^{*\prime}{}_i \mathbf{X}_i^* \left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau | \mathbf{X}_i) - Q_{\mathbf{Y}|\mathbf{X}}(\tau | \mathbf{X}_i)\right) \mathbf{1}(|D_i| < h_N) \xrightarrow{d} Z_1(\tau)$$

with $\mathbb{E}[Z_1(\tau) Z_1(\tau')'] = \pi_0 \mathbb{E}[\mathbf{W}^{*\prime} \mathbf{X}^* \Sigma(\tau, \tau', \mathbf{X}, \mathbf{X}) \mathbf{X}^{*\prime} \mathbf{W}^* | D = 0]$.

Proof.

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_{i}\mathbf{X}_{i}^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{i}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{i}\right)\right)\mathbf{1}(|D_{i}|<h_{N})$$

$$\begin{aligned}
&=\frac{1}{\sqrt{Nh_{N}}}\sum_{i=1}^{N}\mathbf{W}^{*\prime}{}_{i}\mathbf{X}_{i}^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{i}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{i}\right)\right)\mathbf{1}(D_{i}=0)\\
&=\sqrt{N}\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{x}_{l}^{*}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{l}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{l}\right)\right)\hat{p}_{l}\\
&=\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{x}_{l}^{*}\sqrt{Np_{l}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{l}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{l}\right)\right)\frac{\hat{p}_{l}}{\sqrt{p_{l}}}\\
&=\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{x}_{l}^{*}\sqrt{Np_{l}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{l}\right)-Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_{l}\right)\right)\frac{\hat{p}_{l}}{\sqrt{p_{l}\pi_{0}}}\sqrt{\pi_{0}}\\
&\xrightarrow{d}\sqrt{\pi_{0}}\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{x}_{l}^{*}Z_{Q}(\tau,\mathbf{x}_{l})\sqrt{q_{l|0}}\\
&=Z_{1}(\tau)
\end{aligned}$$

Since $E[Z_Q(\tau,\mathbf{x}_l)Z_Q(\tau',\mathbf{x}_m)']=\Sigma(\tau,\tau',\mathbf{x}_l,\mathbf{x}_m)=(\min(\tau,\tau')-\tau\tau')\Lambda(\tau,\tau';\mathbf{x}_l)\cdot\mathbf{1}(l=m)$, we get that:

$$\begin{aligned}
\mathbb{E}\left[Z_{1}(\tau)Z_{1}(\tau')\right]&=\pi_{0}\sum_{l=L+1}^{M}\mathbf{w}^{*\prime}{}_{l}\mathbf{x}_{l}^{*}\Sigma(\tau,\tau',\mathbf{x}_{l},\mathbf{x}_{l})\mathbf{x}_{l}^{*\prime}\mathbf{w}^{*}{}_{l}q_{l|0}\\
&=\pi_{0}\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^{*}\Sigma(\tau,\tau',\mathbf{X},\mathbf{X})\mathbf{X}^{*\prime}\mathbf{W}^{*}|D=0\right].
\end{aligned}$$

■

Proof of proposition 2.3.5.

We complete the proof in the discrete bandwidth case first: $b>0$ and $\pi_0=0$. In this case, $\mathbf{1}(|D_i|<h_N)=\mathbf{1}(D_i=0)$, since we have a (shrinking) mass of "exact" stayers, and no units in between the near-movers ($|D_i|=h_N$) and these stayers. Combining lemmas 2.7.1 and 2.7.2, we can derive the asymptotic distribution of $\widehat{\delta}(\tau)$ in the bandwidth case.

We can use lemmas 2.7.3 and 2.7.4 to prove the analogous result in the discrete movers' case. ■

Proof of propositions 2.4.1.

$$\begin{aligned}\widehat{\bar{\beta}}_I^M(\tau) - \bar{\beta}^M(\tau) = & \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i) - \widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\right)\mathbf{1}(\mathbf{X}_i \in \mathbb{X}^M)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in \mathbb{X}^M)} + \\ & \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1} Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\,\mathbf{1}(\mathbf{X}_i \in \mathbb{X}^M)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i \in \mathbb{X}^M)} - \mathbb{E}\left[\mathbf{X}_i^{-1} Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\,|\mathbf{X}_i \in \mathbb{X}^M\right] \\ & = \sum_{l=1}^{L_1} \mathbf{x}_l^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l) - Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l)\right)\widehat{q}_l^M + \\ & \sum_{l=1}^{L_1} \mathbf{x}_l^{-1} Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l)\,(\widehat{q}_l^M - q_l^M) \\ & = T_1(\tau) + T_2(\tau)\end{aligned}$$

where $q_l^M = \frac{p_l}{1-\pi_0}$, is the conditional probability of realization l conditional on being a mover realization, and let $\widehat{q}_l^M = \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\mathbf{X}_i = \mathbf{x}_l)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(\det(\mathbf{X}_i)\neq 0)}$.

Using Slutsky's theorem, we can see that $\sqrt{N}T_1(\tau) \xrightarrow{d} \sum_{l=1}^{L_1} \mathbf{x}_l^{-1} Z_Q(\tau, \mathbf{x}_l)\frac{\sqrt{p_l}}{1-\pi_0}$, and its asymptotic covariance function is equal to:

$$\begin{aligned}\mathbb{E}\left[\sum_{l=1}^{L_1} \mathbf{x}_l^{-1} Z_Q(\tau, \mathbf{x}_l)\frac{\sqrt{p_l}}{1-\pi_0}\sum_{l=1}^{L_1} Z_Q(\tau', \mathbf{x}_l)'\mathbf{x}_l^{-1\prime}\frac{\sqrt{p_l}}{1-\pi_0}\right] = & \frac{\min(\tau,\tau') - \tau\tau'}{1-\pi_0} \times \\ & \mathbb{E}\left[\mathbf{X}^{-1}\Lambda(\tau,\tau',\mathbf{X})\mathbf{X}^{-1\prime}|X \in \mathbb{X}^M\right].\end{aligned}$$

The second source of variation comes from term $T_2(\tau) = \sum_{l=1}^{L_1} \mathbf{x}_l^{-1} Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l)\,(\widehat{q}_l^M - q_l^M)$. Using the delta method, we can see that this term converges in distribution like such:

$$\sqrt{N}T_2(\tau) \xrightarrow{d} N\left(0, \frac{\text{Var}\left[\beta(\tau, \mathbf{X}|\mathbf{X} \in \mathbb{X}^M)\right]}{1-\pi_0}\right).$$

To conclude, we note that $T_1(\tau)$ and $T_2(\tau)$ are independent, since the source of variation in $T_2(\tau)$ comes solely from variation in $\mathbf{X}$, while the variation in $T_1(\tau)$ is conditional on $\mathbf{X}$, and therefore independent of $\mathbf{X}$. ■

Proof of proposition 2.4.3.

The population estimand is:

$$\begin{aligned}\bar{\beta}_N(\tau) &= \mathbb{E}_N\left[\beta(\tau, \mathbf{X})\right] \\ &= \mathbb{E}_N\left[\beta_D(\tau, D)\right]\end{aligned}$$

where $\beta_D(\tau, D) = \mathbb{E}_N\left[\beta(\tau, \mathbf{X})|D\right]$ and the subscript N on the expectation operator reflects the fact that the support of $\mathbf{X}$ can vary with the sample size. Let's consider a four-term decomposition of this estimator minus its probability limit:

$$\begin{aligned}\widehat{\bar{\beta}}_N(\tau) - \bar{\beta}_N(\tau) = &\frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1}\mathbf{W}_i\mathbf{1}(|D_i| \geq h_N)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| \geq h_N)}\left(\widehat{\delta}(\tau) - \delta(\tau)\right) \\ &+ \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right) - Q_{\mathbf{Y}|\mathbf{X}}\left(\tau|\,\mathbf{X}_i\right)\right)\mathbf{1}(|D_i| \geq h_N)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| \geq h_N)} \\ &+ \frac{\frac{1}{N}\sum_{i=1}^N \beta(\tau, \mathbf{X}_i)\mathbf{1}(|D_i| \geq h_N)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| \geq h_N)} - \frac{\mathbb{E}_N\left[\beta(\tau, \mathbf{X}_i)\mathbf{1}(|D| \geq h_N)\right]}{P_N(|D_i| \geq h_N)} \\ &+ \frac{\mathbb{E}_N\left[\beta(\tau, \mathbf{X})\mathbf{1}(|D| \geq h_N)\right]}{P_N(|D| \geq h_N)} - \mathbb{E}_N\left[\beta(\tau, \mathbf{X})\right] \\ &= T_1(\tau) + T_2(\tau) + T_3(\tau) + T_4(\tau).\end{aligned}$$

We will consider the asymptotic behavior of these four terms separately from $T_4(\tau)$ to $T_1(\tau)$. First, term $T_4(\tau)$ can be shown to be of order $O(h_N)$.

First, we have $P_N(|D| \geq h_N) = 1 - P_N(D = 0) = 1 - 2bh_N$. Also,

$$\begin{aligned}\mathbb{E}_N\left[\beta(\tau, \mathbf{X})\mathbf{1}(|D| \geq h_N)\right] &= \mathbb{E}_N\left[\beta(\tau, \mathbf{X})\right] - \mathbb{E}_N\left[\beta(\tau, \mathbf{X})\mathbf{1}(|D| = 0)\right] \\ &= \mathbb{E}_N\left[\beta(\tau, \mathbf{X})\right] - E_N\left[\beta(\tau, \mathbf{X})|D = 0\right]P_N(D = 0) \\ &= \mathbb{E}_N\left[\beta(\tau, \mathbf{X})\right] - \beta_D(\tau, 0)2bh_N.\end{aligned}$$

Therefore, we have that:

$$
\begin{aligned}
T_4(\tau) &= \frac{\mathbb{E}_N\left[\beta(\tau,\mathbf{X})\mathbf{1}(|D|\geq h_N)\right]}{P_N(|D|\geq h_N)} - \mathbb{E}_N\left[\beta(\tau,\mathbf{X})\right] \\
&= \frac{\left(\mathbb{E}_N\left[\beta(\tau,\mathbf{X})\right] - \beta_D(\tau,0)\right)2bh_N}{1-2bh_N} \\
&= O(h_N).
\end{aligned}
$$

Then, term $T_3(\tau)$ can be shown to be of order $O_p\left(\frac{1}{\sqrt{N}}\right)$. We will use the delta method to find the asymptotic order of the distribution of term $T_3(\tau)$. First, let

$$
Z_{N,i} = \begin{pmatrix} \beta(\tau,\mathbf{X}_i)\mathbf{1}(|D_i|\geq h_N) - \mathbb{E}_N\left[\beta(\tau,\mathbf{X})\mathbf{1}(|D|\geq h_N)\right] \\ \mathbf{1}(|D_i|\geq h_N) - P_N(|D|\geq h_N) \end{pmatrix}.
$$

Then, the variance of $Z_{N,i}$ is equal to:

$$
\begin{aligned}
\mathrm{Var}\left[Z_{N,i}\right] &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix} \\
\Sigma_{11} &= \mathbb{V}_N(\beta(\tau,\mathbf{X})) - E\left[\beta(\tau,\mathbf{X})\beta(\tau,\mathbf{X})'\mathbf{1}(D=0)\right] + \\
&\quad \beta_D(\tau,0)2bh_N(2\mathbb{E}_N\left[\beta(\tau,\mathbf{X})\mathbf{1}(|D|\geq h_N)\right] - \beta_D(\tau,0)2bh_N) \\
&= \mathbb{V}_N(\beta(\tau,\mathbf{X})) + O(h_N) \\
\Sigma_{12} &= 2bh_N(\mathbb{E}_N\left[\beta(\tau,\mathbf{X})\right] - \beta_D(\tau,0)) \\
\Sigma_{22} &= (1-2bh_N)2bh_N
\end{aligned}
$$

Using these results and Lyapunov's CLT, we see that $\frac{1}{N}\sum_{i=1}^N Z_{N,i}$ converges in distribution if premultiplied by a matrix $\Psi_N = \begin{pmatrix} \sqrt{N} & 0 \\ 0 & \sqrt{\frac{N}{h_N}} \end{pmatrix}$. Using the delta method, we see that

$$
\sqrt{N}\left(\frac{\frac{1}{N}\sum_{i=1}^N \beta(\tau,\mathbf{X}_i)\mathbf{1}(|D_i|\geq h_N)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i|\geq h_N)} - \frac{\mathbb{E}_N\left[\beta(\tau,\mathbf{X})\mathbf{1}(|D|\geq h_N)\right]}{P_N(|D|\geq h_N)}\right) \xrightarrow{d} N\left(0, \lim_{N\to\infty} \mathbb{V}_N(\beta(\tau,\mathbf{X}))\right),
$$

since the denominator converges in distribution faster than the numerator. We now check that term $T_2(\tau)$ will be of order $O_p\left(\frac{1}{\sqrt{Nh_N}}\right)$. First, we see that the denominator

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(|D_i| \geq h_N) \xrightarrow{\mathbb{P}} 1$$

if $h_N \to 0$ as $N \to \infty$.

The numerator, $\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i) - Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\right)\mathbf{1}(|D_i| \geq h_N)$ contains units from both strict-movers ($|D_i| \nrightarrow 0$) and for near-movers ($|D_i| > 0$ and $D_i \to 0$). We can linearly decompose the numerator into two terms, one of which containing near-stayers and one for strict-movers.

Consider first the term $T_2(\tau)'$ which contains the strict-movers:

$$\begin{aligned}
\sqrt{N}T_2(\tau)' &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{X}_i^{-1}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i) - Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\right)\mathbf{1}(|D_i| > h_N) \\
&= \sum_{l=1}^{L_1}\mathbf{x}_l^{-1}\sqrt{Np_{lN}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l) - Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_l)\right)\frac{\hat{p}_{lN}}{\sqrt{p_{lN}}} \\
&\xrightarrow{d} \sum_{l=1}^{L_1}\mathbf{x}_l^{-1}Z_Q(\tau,\mathbf{x}_l)\sqrt{p_l}.
\end{aligned}$$

Consider first the term $T_2(\tau)''$ which contains the near-stayers:

$$\begin{aligned}
\sqrt{Nh_N}T_2(\tau)'' &= \sqrt{Nh_N}\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{X}_i^*}{D_i}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i) - Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{X}_i)\right)\mathbf{1}(|D_i| = h_N) \\
&= \sum_{l=L_1+1}^{L}\mathbf{x}_l^*\sqrt{Np_{lN}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_{lN}) - Q_{\mathbf{Y}|\mathbf{X}}(\tau|\,\mathbf{x}_{lN})\right)\frac{\hat{p}_{lN}}{\sqrt{p_{lN}2bh_N}}\sqrt{2b} \\
&\xrightarrow{d} \sum_{l=L_1+1}^{L}\mathbf{x}_l^*Z_Q(\tau,\mathbf{x}_l)\sqrt{q_{l|0}2b}.
\end{aligned}$$

Now, finally, we consider term $T_1(\tau)$. $\widehat{\delta}(\tau) - \delta(\tau)$ is of order $O_p\left(\frac{1}{\sqrt{Nh_N}}\right)$ and is independent of term $T_2(\tau)$, since it is computed with a different segment of the sample. We also

consider the probability limit of the term attached to it, $\widehat{\widehat{\Xi}}_N = \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{X}_i^{-1}\mathbf{W}_i\mathbf{1}(|D_i| \geq h_N)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| \geq h_N)}$:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \frac{\mathbf{W}_i^*\mathbf{1}(|D_i| \geq h_N)}{D_i} &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{W}^*|D = h_N\right]\frac{1}{h_N}\mathbf{1}(D_i = h_N)\right] - \\
&\quad \mathbb{E}\left[\mathbb{E}\left[\mathbf{W}^*|D = -h_N\right]\frac{1}{h_N}\mathbf{1}(D_i = -h_N)\right] + \\
&\quad \mathbb{E}\left[\frac{\mathbf{W}_i^*\mathbf{1}(|D_i| > h_N)}{D_i}\right] + o_P(1) \\
&= b(\mathbb{E}\left[\mathbf{W}^*|D = h_N\right] - \mathbb{E}\left[\mathbf{W}^*|D = -h_N\right]) + O_p(1) \\
&= O_p(h_N) + O_p(1)
\end{aligned}$$

We can then see that $\widehat{\widehat{\Xi}}_N$ has a well defined probability limit. We can now compute the asymptotic distribution of $\widehat{\widehat{\beta}}^M(\tau) - \bar{\beta}(\tau)$ easily. Since $Nh_N^3 \to 0$, the bias term vanishes, and and so does the term containing the variance of $\beta(\tau; \mathbf{X}_i)$, since it will be of order root-h_N when multiplied by the rate $\sqrt{Nh_N}$. This completes the proof. ■

Proof of proposition 2.5.1.

We start by deriving the asymptotic distribution of the CDF of $\widehat{\beta}(U; \mathbf{X})$ with U unformly $[0, 1]$ distributed, independently from $\mathbf{X}$:

$$\begin{aligned}
\sum_{l=1}^L \int_0^1 \mathbf{1}(\widehat{\beta}_p(u, \mathbf{x}_l) \leq c)du\widehat{q}_l^M - F_{\mathbf{B}_p|\mathbf{X}\in\mathbb{X}^M}(c) &= \sum_{l=1}^L \int_0^1 \mathbf{1}(\widehat{\beta}_p(u, \mathbf{x}_l) \leq c)du\widehat{q}_l^M - \\
&\quad \sum_{l=1}^L \int_0^1 \mathbf{1}(\beta_p(u, \mathbf{x}_l) \leq c)duq_l^M \\
&= \sum_{l=1}^L \left(\int_0^1 \mathbf{1}(\widehat{\beta}_p(u, \mathbf{x}_l) \leq c)du\widehat{q}_l^M - \int_0^1 \mathbf{1}(\beta_p(u, \mathbf{x}_l) \leq c)du\right)\widehat{q}_l^M + \\
&\quad + \sum_{l=1}^L \int_0^1 \mathbf{1}(\beta_p(u, \mathbf{x}_l) \leq c)du\left(\widehat{q}_l^M - q_l^M\right) \\
&= T_1(c) + T_2(c)
\end{aligned}$$

Using the functional delta method on term $T_1(\tau)$, we can see that:

$$\begin{aligned}
\sqrt{N}T_1(c) &= \sqrt{N}\sum_{l=1}^{L}\left(\int_0^1 \mathbf{1}(\widehat{\beta}_p(u,\mathbf{x}_l)\leq c)du - \int_0^1 \mathbf{1}(\beta_p(u,\mathbf{x}_l)\leq c)du\right)\widehat{q}_l^M \\
&\xrightarrow{d} Z_1(c) \\
\mathbb{E}\left[Z_1(c)Z_1(c')'\right] &= \mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X})\mathbf{X}^{-1}(\min(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X}),F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X}))-\right. \\
&\quad F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X})F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X}))\Lambda(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X}),F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X}),\mathbf{X})\times \\
&\quad \left.\mathbf{X}^{-1\prime}f_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X})|\mathbf{X}\in\mathbb{X}^M\right]+ \\
&\quad \frac{1}{\pi_0}\mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X})\mathbf{X}^{-1}\mathbf{W}Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X}))\times\right. \\
&\quad \left.Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(c',\tilde{\mathbf{X}}))'\tilde{\mathbf{W}}'\tilde{\mathbf{X}}^{-1\prime}f_{\mathbf{B}_p|\mathbf{X}}(c',\tilde{\mathbf{X}})|\mathbf{X},\tilde{\mathbf{X}}\in\mathbb{X}^M\right]
\end{aligned}$$

The second term's convergence is straightforward and similar to previous proofs:

$$\begin{aligned}
\sqrt{N}T_2(c) &= \sqrt{N}\sum_{l=1}^{L}\int_0^1 \mathbf{1}(\beta_p(u,\mathbf{x}_l)\leq c)du(\widehat{q}_l^M - q_l^M) \\
&\xrightarrow{d} Z_2(c) \\
\mathbb{E}\left[Z_2(c)Z_2(c')'\right] &= \frac{1}{1-\pi_0}\mathrm{Cov}\left(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X}),F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X})|\mathbf{X}\in\mathbb{X}^M\right)
\end{aligned}$$

Let $Z_F(c) = Z_1(c) + Z_2(c)$ and observe that $Z_1(\cdot)$ and $Z_2(\cdot)$ are independent since they are computed with different subsamples. Then, using the delta method, we get that

$$\sqrt{N}\left(\widehat{\beta}_p^M(\tau) - \beta_p^M(\tau)\right) \xrightarrow{d} \frac{Z_F(\beta_p^M(\tau))}{f_{\mathbf{B}_p|\mathbf{X}\in\mathbb{X}^M}(\beta_p^M(\tau))}.$$

■

Proof of proposition 2.5.2.

We again start by deriving the asymptotic distribution of the CDF of $\widehat{\beta}(U;\mathbf{X})$ with U independent from $\mathbf{X}$:

$$\sum_{l=1}^{L}\int_0^1 \mathbf{1}(\widehat{\beta}_{pN}(u,\mathbf{x}_l)\leq c)du\widehat{q}_{lN}^M - F_{\mathbf{B}_p}(c) = \sum_{l=1}^{L}\int_0^1 \mathbf{1}(\widehat{\beta}_{pN}(u,\mathbf{x}_{lN})\leq c)du\widehat{q}_{lN}^M -$$
$$\sum_{l=1}^{L}\int_0^1 \mathbf{1}(\beta_{p}(u,\mathbf{x}_{lN})\leq c)duq_{lN}^M$$
$$= \sum_{l=1}^{L}\left(\int_0^1 \mathbf{1}(\widehat{\beta}_{p}(u,\mathbf{x}_{lN})\leq c)du - \int_0^1 \mathbf{1}(\beta_{p}(u,\mathbf{x}_{lN})\leq c)du\right)\widehat{q}_{lN}^M +$$
$$\sum_{l=1}^{L}\int_0^1 \mathbf{1}(\beta_{p}(u,\mathbf{x}_{lN})\leq c)du\left(\widehat{q}_{lN}^M - q_{lN}^M\right) +$$
$$F_{\mathbf{B}_p|\mathbf{X}\in\mathbb{X}_N^M}(c) - F_{\mathbf{B}_p}(c)$$
$$= T_1(c) + T_2(c) + T_3(c)$$

We can decompose the mover realizations in $T_1(c)$ into its near-stayer realizations ($T_1''(c)$) and its strict mover realizations ($T_1'(c)$):

$$T_1(c) = T_1'(c) + T_1''(c)$$
$$T_1'(c) = \sum_{l=1}^{L_1}\left(\int_0^1 \mathbf{1}(\widehat{\beta}_{pN}(u,\mathbf{x}_{lN})\leq c)du - \sum_{l=1}^{L_1}\int_0^1 \mathbf{1}(\beta_{p}(u,\mathbf{x}_{lN})\leq c)du\right)\widehat{q}_{lN}^M$$
$$= O_p\left(\frac{1}{\sqrt{N}}\right)$$

because the conditional betas are root-N consistent for strict movers. In the case of near-stayers, $T_1''(c)$ will be root-Nh_N consistent because:

$$\sqrt{Nh_N}T_1''(c) = \sum_{l=L_1+1}^{L}\sqrt{Nh_N^3}\left(\int_0^1 \mathbf{1}(\widehat{\beta}_{pN}(u,\mathbf{x}_{lN})\leq c)du - \int_0^1 \mathbf{1}(\beta_{p}(u,\mathbf{x}_{lN})\leq c)du\right)\frac{2b\widehat{q}_{lN}^M}{2bh_N}$$

and since the conditional betas for near-stayers are of order $O_p\left(\frac{1}{\sqrt{Nh_N^3}}\right)$, this expression will converge in distribution. The probabilities $\frac{\widehat{q}_{lN}^M}{2bh_N}$ will converge in probability to $q_{l|0}$. Using

the functional delta method, we can get that:

$$\sqrt{Nh_N^3}\left(\int_0^1 \mathbf{1}(\widehat{\beta}_{pN}(u,\mathbf{x}_{lN})\leq c)du - \int_0^1 \mathbf{1}(\beta_p(u,\mathbf{x}_{lN})\leq c)du\right) \xrightarrow{d} Z'(l,c)$$

$$\begin{aligned}\mathbb{E}\left[Z'(l,c)Z'(l',c')'\right] = \frac{1}{2bq_{l|0}} f_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{x}_l) f_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{x}_l)\mathbf{x}_l^* \big(\min(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{x}_l), F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{x}_l)) - \times \\ F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{x}_l)F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{x}_l)\big)\,\Lambda(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{x}_l), F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{x}_l),\mathbf{X})\mathbf{x}_l^{*\prime}\mathbf{1}(l=l') + \\ \frac{1}{2b} f_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{x}_l)\mathbf{w}_l^*\mathbb{E}\left[Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{x}_l))Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{x}_{l'}))'\right]\mathbf{w}_{l'}^{*\prime} f_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{x}_{l'}).\end{aligned}$$

Using the Slutsky's theorem, we get that

$$\begin{aligned}&\sqrt{Nh_N}T_1''(c) \xrightarrow{d} Z_1(c)\\ &\mathbb{E}\left[Z_1(c)Z_1(c')'\right] = 4b^2\sum_{l=L_1+1}^{L}\sum_{l'=L_1+1}^{L}\mathbb{E}\left[Z'(l,c)Z'(l',c')'\right]q_{l|0}q_{l'|0}\\ &\quad= 2b\mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X})\mathbf{X}^*(\min(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X}), F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X}) - F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X})F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X}))\times\right.\\ &\qquad\left.\Lambda(F_{\mathbf{B}_p|\mathbf{X}}(c,\mathbf{X}), F_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X}),\mathbf{X})\mathbf{X}^{*\prime}f_{\mathbf{B}_p|\mathbf{X}}(c',\mathbf{X})|D=0\right] +\\ &\quad 2b\mathbb{E}\left[f_{\mathbf{B}_p|\mathbf{X}}(d,\mathbf{X})\mathbf{W}^* Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau),\mathbf{X}))\right.\\ &\qquad\left.Z_\delta(F_{\mathbf{B}_p|\mathbf{X}}(\beta_p(\tau'),\tilde{\mathbf{X}}))'\tilde{\mathbf{W}}^{*\prime}f_{\mathbf{B}_p|\mathbf{X}}(d',\tilde{\mathbf{X}})|D=0,\tilde{D}=0\right].\end{aligned}$$

Using similar arguments to the proof for the ACQE, we can see that $T_2(c)$ is of order $O_p\left(\frac{1}{\sqrt{N}}\right)$, and therefore will be asymptotically negligible. Finally, $T_3(c)$ will be of order $O(h_N)$ with a similar argument as well. This, along with the assumption that $Nh_N^3 \to 0$ will imply that $Z_1(c)$ is the only asymptotic component in the estimation of the unconditional CDF of the random coefficient. Using this fact, we can see that:

$$Z_{UQE}(\tau) = \frac{Z_1(\beta_p(\tau))}{f_{\mathbf{B}_p}(\beta_p(\tau))}.$$

■

Bibliography

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, 59(2):307–45.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Belloni, A., Chernozhukov, V., and Fernandez-Val, I. (2011). Conditional quantile processes based on series or many regressors. *arXiv preprint arXiv:1105.6154*.

Berry, S., Gandhi, A., and Haile, P. (2012). Connected substitutes and invertibility of demand. Mimeo, Yale University.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):pp. 841–890.

Bester, C. A. and Hansen, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics*, 27(2):235–250.

Bhattacharya, P. K. and Zhao, P. L. (1997). Semiparametric inference in a partial linear model. *Annals of Statistics*, 25:244–262.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Springer.

Brown, B. W. and Newey, W. K. (1998). Efficient semiparametric estimation of expectations. *Econometrica*, 66(2):453–464.

Brown, D. J. and Wegkamp, M. H. (2002). Weighted minimum mean-square distance from independence estimation. *Econometrica*, 70(5):2035–2051.

Carrasco, M. and Florens, J.-P. (2000). Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16:797–834.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46.

Chamberlain, G. (1984). Panel data. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume 2 of *Handbook of Econometrics*, chapter 22, pages 1247–1318. Elsevier.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10(1):pp. 20–26.

Chaudhuri, P., Doksum, K., and Samarov, A. (1997). On average derivative quantile regression. *The Annals of Statistics*, 25(2):715–744.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2009). Inference on counterfactual distributions.

Chernozhukov, V., Fernndez-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, 71(5):1405–1441.

Cosslett, S. (1987). Efficiency bounds for distribution free estimators of the binary choice model. *Econometrica*, 55(3):559–585.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B*, 41:1–31.

Domínguez, M. A. and Lobato, I. N. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72(5):1601–1615.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117:55–93.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2008). Choosing the number of moments in conditional moment restriction models. Mimeo, University of Texas.

Evdokimov, K. (2009). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. manuscript, Department of Economics, Yale University.

Gallant, A. R. and Souza, G. (1991). On the asymptotic normality of fourier flexible form estimates. *Journal of Econometrics*, 50(3):329–353.

Graham, B. S. and Powell, J. L. (2012). Identification and estimation of average partial effects in irregular correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model : The maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316.

Hansen, C., McDonald, J. B., and Newey, W. K. (2010). Instrumental variables estimation with flexible distributions. *Journal of Business and Economic Statistics*, 28(1):13–25.

Hsieh, D. A. and Manski, C. F. (1987). Monte carlo evidence on adaptive maximum likelihood estimation of a regression. *The Annals of Statistics*, 15(2):541–551.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(12):71 – 120.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Komunjer, I. and Santos, A. (2010). Semi-parametric estimation of non-separable models: A minimum distance from independence approach. *Econometrics Journal*, 13:S28–S55.

Lee, L.-F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics*, 65(2):381–428.

Lemieux, T. (2006). Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *The American Economic Review*, pages 461–498.

Linton, O. and Gozalo, P. (1996). Conditional independence restrictions: Testing and estimation. Mimeo, Yale University and Brown University.

Lukacs, E. (1960). *Characteristic Functions*. Griffin's Statistical Monographs and Courses. Charles Griffin & Company Limited.

MaCurdy, T. E. (1982). Using information on the moments of disturbances to increase the efficiency of estimation. Working Paper 22, National Bureau of Economic Research.

Manski, C. F. (1983). Closest empirical distribution estimation. *Econometrica*, 51(2):305–319.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55(2):357–62.

Melly, B. (2006). Estimation of counterfactual distributions using quantile regression. *Review of Labor Economics*, 68:543–572.

Newey, W. K. (1989). Locally efficient, residual-based estimation of nonlinear simultaneous equations models. Mimeo, Princeton University.

Newey, W. K. (1990a). Efficient estimation of semiparametric models via moment restrictions. Mimeo, Princeton University.

Newey, W. K. (1990b). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–37.

Newey, W. K. (1990c). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. volume 11, chapter 16.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, chapter 36, pages 2111–2245. Elsevier.

Parzen, E. (1959). Statistical inference on time series by hilbert space methods, i. Technical Report 23, Stanford University.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303 – 325.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57(6):pp. 1403–1430.

Qu, Z. and Yoon, J. (2011). Nonparametric estimation and inference on conditional quantile processes. Technical report, Boston University-Department of Economics.

Robinson, P. M. (1988). Root- n-consistent semiparametric regression. *Econometrica*, 56(4):931–54.

Rothe, C. (2010). Decomposing counterfactual distributions. Technical report, Toulouse School of Economics.

Ruud, P. A. (2000). Semiparametric estimation of discrete choice models. manuscript, Department of Economics, University of California at Berkeley.

Santos, A. (2011). Semiparametric estimation of invertible models. Mimeo, University of California, San Diego.

Su, L. and White, H. (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864.

Thompson, T. S. (1993). Some efficiency bounds for semiparametric discrete choice models. *Journal of Econometrics*, 58:257–274.

Torgovitsky, A. (2012). Identification of nonseparable models with general instruments. Mimeo, Northwestern University.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer.