ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)

MALMEM: model averaging in linear measurement error models

Xinyu Zhang,

University of Science and Technology of China, Hefei, and Chinese Academy of Sciences, Beijing, People's Republic of China

Yanyuan Ma

Pennsylvania State University, University Park, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA, and University of Technology Sydney, Australia

[Received January 2018. Revised February 2019]

Summary. We develop model averaging estimation in the linear regression model where some covariates are subject to measurement error. The absence of the true covariates in this framework makes the calculation of the standard residual-based loss function impossible. We take advantage of the explicit form of the parameter estimators and construct a weight choice criterion. It is asymptotically equivalent to the unknown model average estimator minimizing the loss function. When the true model is not included in the set of candidate models, the method achieves optimality in terms of minimizing the relative loss, whereas, when the true model is included, the method estimates the model parameter with root *n* rate. Simulation results in comparison with existing Bayesian information criterion and Akaike information criterion model selection and model averaging methods strongly favour our model averaging method. The method is applied to a study on health.

Keywords: Measurement error; Model averaging; Model selection; Optimality; Weight

1. Introduction

Many data sets in real life contain measurement error. For example, in nutrition studies, food intake measurements rely on self-reported consumption through food questionnaires, recalls or diaries. In biomedical studies, biomarkers are measured from assays and can contain substantial error due to human effect or laboratory conditions. Descriptions of various measurement error problems and their treatments are available for both linear models (Fuller, 1987) and non-linear models (Buonaccorsi, 2010; Carroll *et al.*, 2006; Gustafson, 2004) in the statistics literature. Similarly to the case when covariates are precisely measured, when studying a data set with covariates measured with errors, practitioners often have many candidate models and model selection methods are generally utilized to select the most suitable model.

Model averaging is an alternative to model selection. When model selection is used, the implicit

Address for correspondence: Xinyu Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, East Zhong-Guan-Cun Road, Beijing 100190, People's Republic of China.
E-mail: xinyu@amss.ac.cn

1369–7412/19/81000

assumption is that one model is ‘correct’ or is at least ‘more correct’ than all others. In reality, however, it can happen that all the models under consideration are wrong, but several competitive models are equally or similarly suitable for the data at hand. For example, when we use a model selection criterion to choose a model, several models may yield very close criterion values. This indicates that no single model obviously dominates all other models. In this case, using a single model may impose some risk in the subsequent analysis, as we are ‘putting all our inferential eggs in one unevenly woven basket’ (Longford, 2005). Even when there is a single model which obviously dominates all other models, the probability of choosing this model via a criterion is generally smaller than 1, because sample size is finite in practice. In this case, when a wrong model is selected, the subsequent analysis will be invalid. Because of these considerations, compared with model selection, model averaging has its advantage. It combines models instead of choosing a single one of them and can be considered as a more prudent way of proceeding with data modelling.

Model averaging has long been a popular approach within the Bayesian paradigm; see, for example, Hoeting *et al.* (1999) for a comprehensive review. In recent years, frequentist model averaging has also been actively developed. Buckland *et al.* (1997) suggested a general approach of assigning model weights based on the scores of information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). This weighting strategy was also used by Hjort and Claeskens (2003), Zhang and Liang (2011) and Zhang *et al.* (2012). Hansen (2007), a seminal work on asymptotically optimal model averaging, selected the weights through minimizing the Mallows criterion, because of its unbiasedness (up to a constant) in estimating expected squared error. Other frequentist model averaging strategies include adaptive regression through mixing (Yang, 2001), jackknife model averaging (Hansen and Racine, 2012), heteroscedasticity robust model averaging (Liu and Okui, 2013), model averaging marginal regression (Chen *et al.*, 2018; Li *et al.*, 2015) and the plug-in method (Liu, 2015). Model averaging has also been extended to other contexts such as structural break models (Hansen, 2009), mixed effects models (Zhang *et al.*, 2014), factor-augmented regression models (Cheng and Hansen, 2015), quantile regression models (Lu and Su, 2015), generalized linear models (Zhang *et al.*, 2016) and missing data models (Fang *et al.*, 2019; Zhang, 2013).

When covariates are measured with error, we face the same problems about model selection. Thus it is natural to opt for model averaging and to study how to choose model averaging weights. However, studies regarding weight choice for model averaging when covariates are measured with errors are essentially non-existent. In fact, the only work that is related to model averaging in measurement error models is Wang *et al.* (2012), where inference after model averaging was studied, but no weight choice method was proposed. One fundamental difficulty in performing model averaging for measurement error problems is that residuals cannot be formed when the true covariates are unavailable, regardless of how well the parameters are estimated in any given model. In addition, likelihoods or even the observed data distribution functions are also unavailable or not computable in measurement error problems. As a consequence, none of the existing asymptotically optimal model averaging methods such as weight choices based on Mallows and jackknife criteria applies. Although the criterion-based model average methods such as the smoothed AIC (SAIC) or smoothed BIC (SBIC) (Buckland *et al.*, 1997) could be applied, these are *ad hoc* approaches in the measurement error context and their properties are not known. This motivates us to fill this literature gap and to initiate researches in model averaging under covariate measurement error. We study how best to average different linear measurement error models through choosing model weights in a data-driven fashion via fully exploiting the inherent properties of the model. The resulting model averaging estimator is asymptotically optimal in the sense that it is asymptotically equivalent to the optimal but infeasible model average estimator that minimizes the loss function. This result is useful in prediction when future observation becomes available which no longer involves measurement error (Carroll *et al.*, 2009), as is the case in the data example that is illustrated in Section 4, where a validation data set without measurement error is available. We also numerically illustrate that the proposed model averaging method is superior to commonly used model averaging and selection methods. We emphasize that, in the simpler case where the same measurement error structure is retained in the available data as well as in any future data where prediction is to be conducted, there is no real need to take into account the measurement error issues (Buonaccorsi, 2010; Carroll *et al.*, 2006).

The paper is organized as follows. In Section 2, we describe the model framework, propose a weight choice criterion and show the asymptotic properties of the resulting model averaging estimator. We conduct simulation studies in Section 3 to illustrate the numerical performance of our method and apply our method to a study of health in Section 4. We finish with some discussion in Section 5. All the proofs and technical details are in the on-line supplementary material.

2. Estimation by model averaging

2.1. Model and estimators

Consider the data-generating process

$$Y_i = \mu_{0i} + \epsilon_i, \tag{1}$$

where Y_i is a univariate response, μ_{0i} is the mean of Y_i and the error ϵ_i has mean 0 and variance σ^2. Let $\mathbf{X}_i$ be a p-dimensional covariate vector that is used to predict μ_{0i}. We approximate the relationship between μ_{0i} and $\mathbf{X}_i$ by using a linear model, i.e. $\mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a p-dimensional vector. There is a distinction between μ_{0i} and μ_i, where μ_{0i} denotes the true mean of Y_i and μ_i denotes the mean under the assumed model. Further, some or all components of $\mathbf{X}_i$ are measured with errors. Thus, instead of observing $\mathbf{X}_i$, we observe a p-dimensional random variable $\mathbf{Z}_i$, where $\mathbf{Z}_i = \mathbf{X}_i + \mathbf{U}_i$, and $\mathbf{U}_i$ is independent of $\mathbf{X}_i$ and has a normal distribution with mean 0 and variance–covariance matrix $\boldsymbol{\Sigma}$. To increase flexibility, we allow some components of $\mathbf{U}_i$ to be identically 0; therefore these components of $\mathbf{X}_i$ are precisely measured. This also allows us to include a constant 1 in $\mathbf{X}_i$. Without loss of generality, we shall assume that the last p^* components of $\mathbf{X}_i$ are subject to error, whereas the remaining $p - p^*$ components are error free. Thus, the upper $p - p^*$ subvector of $\mathbf{U}_i$ is zero, and $\boldsymbol{\Sigma}$ is zero except for its lower right-hand $p^* \times p^*$ block. We also assume that the measurement error vector $\mathbf{U}_i$ is independent of ϵ_i, and $(\mathbf{U}_i, \epsilon_i)$ are identically distributed for $i = 1, \ldots, n$.

When taking $\mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, we are in the framework of the well-studied linear measurement error models; see Fuller (1987), Carroll *et al.* (2006) and references therein for a comprehensive review of this literature. Specifically, we can obtain an estimator of $\boldsymbol{\beta}$ through solving $n^{-1}\Sigma_{i=1}^{n}\mathbf{Z}_i(Y_i - \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Sigma}\boldsymbol{\beta} = \mathbf{0}$. This leads to a closed form estimator $\hat{\boldsymbol{\beta}} = (\Sigma_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}} - n\boldsymbol{\Sigma})^{-1}\Sigma_{i=1}^{n}\mathbf{Z}_i\mathbf{Y}_i$. However, in practice, the relationship $\mu_{0i} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ almost never holds for any $\boldsymbol{\beta}$, i.e. $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ is only an approximation of the true regression relationship between $\mathbf{X}_i$ and μ_{0i}. Thus, to alleviate the damage due to the potential model misspecification, we adopt a model averaging approach. The basic idea of model averaging is to use the average of the estimates of a common target quantity from several models, instead of focusing on just one selected specific model. The art of it is in selecting the weights that are associated with the different potential models. In our context, the common target quantity is the mean μ_{0i}.

To explain the central idea of the model averaging estimator better we first treat $\boldsymbol{\Sigma}$ and σ^2 as known. We shall later replace them with their respective estimators $\hat{\boldsymbol{\Sigma}}$ and $\hat{\sigma}^2$ in constructing the weights of our model averaging method, showing that this does not affect model averaging optimality. We shall also prove the asymptotic optimality of our estimator based on $\hat{\boldsymbol{\Sigma}}$ and $\hat{\sigma}^2$. Define $\mathbf{Y}=(Y_1,\ldots,Y_n)^{\mathrm{T}}\in\mathbb{R}^n$, $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^{\mathrm{T}}\in\mathbb{R}^{n\times p}$, $\mathbf{Z}=(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)^{\mathrm{T}}\in\mathbb{R}^{n\times p}$, $\boldsymbol{\mu}_0=(\mu_{01},\ldots,\mu_{0n})^{\mathrm{T}}\in\mathbb{R}^n$, $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^{\mathrm{T}}\in\mathbb{R}^n$ and $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_n)^{\mathrm{T}}\in\mathbb{R}^n$.

Assume that we have a total of S candidate models. In the sth model, we use the candidate model $\boldsymbol{\mu}=\mathbf{X}_{(s)}\boldsymbol{\beta}_{(s)}$ where $\mathbf{X}_{(s)}$ is the $n\times p_s$ regression matrix and $\boldsymbol{\beta}_{(s)}$ is the corresponding coefficient vector; $\mathbf{Z}_{(s)}$, $\mathbf{U}_{(s)}$ and $\boldsymbol{\Sigma}_{(s)}$ are defined similarly. Under this model, the estimator of $\boldsymbol{\beta}_{(s)}$ is $\hat{\boldsymbol{\beta}}_{(s)}=(\mathbf{Z}_{(s)}^{\mathrm{T}}\mathbf{Z}_{(s)}-n\boldsymbol{\Sigma}_{(s)})^{-1}\mathbf{Z}_{(s)}^{\mathrm{T}}\mathbf{Y}$. Let $\mathbf{X}_i^{\mathrm{T}}$ and $\mathbf{X}_{(s),i}^{\mathrm{T}}$ be the ith rows of $\mathbf{X}$ and $\mathbf{X}_{(s)}$ respectively. Let $\boldsymbol{\Pi}_{(s)}$ be the projection matrix mapping $\mathbf{X}_i$ to its subvector $\mathbf{X}_{(s),i}=\boldsymbol{\Pi}_{(s)}\mathbf{X}_i$. Obviously, we also have $\mathbf{X}\boldsymbol{\Pi}_{(s)}^{\mathrm{T}}=\mathbf{X}_{(s)}$. To shorten the notation, let $\mathbf{G}_{(s)}=\boldsymbol{\Pi}_{(s)}^{\mathrm{T}}(\mathbf{Z}_{(s)}^{\mathrm{T}}\mathbf{Z}_{(s)}-n\boldsymbol{\Sigma}_{(s)})^{-1}\mathbf{Z}_{(s)}^{\mathrm{T}}$ and $\mathbf{P}_{(s)}=\mathbf{X}\mathbf{G}_{(s)}$. Then, if we could observe $\mathbf{X}$, the estimator of $\boldsymbol{\mu}_0$ by the sth model based on the measurement error estimator would be

$$\hat{\boldsymbol{\mu}}_{(s)}=\mathbf{X}\boldsymbol{\Pi}_{(s)}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(s)}=\mathbf{X}\mathbf{G}_{(s)}\mathbf{Y}=\mathbf{P}_{(s)}\mathbf{Y}.$$

Let the weight vector be $\mathbf{w}$, where $\mathbf{w}=(w_1,\ldots,w_S)^{\mathrm{T}}$ and it belongs to the set

$$\mathcal{W}=\{\mathbf{w}\in[0,1]^S:\sum_{s=1}^{S}w_s=1\}.$$

The model average estimator of $\boldsymbol{\mu}_0$ would then be

$$\hat{\boldsymbol{\mu}}(\mathbf{w})=\sum_{s=1}^{S}w_s\hat{\boldsymbol{\mu}}_{(s)}=\sum_{s=1}^{S}w_s\mathbf{P}_{(s)}\mathbf{Y}=\mathbf{X}\mathbf{G}(\mathbf{w})\mathbf{Y}=\mathbf{P}(\mathbf{w})\mathbf{Y},$$

where $\mathbf{G}(\mathbf{w})\equiv\Sigma_{s=1}^{S}w_s\mathbf{G}_{(s)}$ and $\mathbf{P}(\mathbf{w})\equiv\Sigma_{s=1}^{S}w_s\mathbf{P}_{(s)}$. We define the squared loss of $\hat{\boldsymbol{\mu}}(\mathbf{w})$ to be $L(\mathbf{w})\equiv\|\hat{\boldsymbol{\mu}}(\mathbf{w})-\boldsymbol{\mu}_0\|^2$, and the risk to be $R(\mathbf{w})\equiv E\{L(\mathbf{w})\}$. To select the optimal weights, we could minimize an approximated version of $R(\mathbf{w})$ with respect to $\mathbf{w}$, if $\mathbf{X}$ had been observed.

Of course $\mathbf{X}$ is not observed. Next, we explain how to construct a criterion $C(\mathbf{w})$ that bypasses $\mathbf{X}$, and at the same time estimates $R(\mathbf{w})$ without bias up to a shift that is unrelated to $\mathbf{w}$. We then minimize $C(\mathbf{w})$ with respect to $\mathbf{w}$, following general model averaging practice (Hansen, 2007; Liang *et al.*, 2011).

2.2. Weight choice criterion

To write out the criterion $C(\mathbf{w})$ explicitly, we first need to introduce some auxiliary quantities. Let $\mathbf{h}_j$ be the jth column of the $p\times p$ identity matrix $\mathbf{I}_p$ and let $\mathbf{b}_i$ be the ith column of $\mathbf{I}_n$. We define $\hat{\boldsymbol{\epsilon}}(\mathbf{w})\equiv\mathbf{Y}-\mathbf{Z}\mathbf{G}(\mathbf{w})\mathbf{Y}$ and $\tilde{\mathbf{U}}_i\equiv\boldsymbol{\Sigma}^{-1/2}\mathbf{U}_i$, where $\boldsymbol{\Sigma}^{-1/2}$ is the matrix whose lower right-hand block is the square root of the inverse of the same block of the matrix $\boldsymbol{\Sigma}$, and the rest of the entries are 0s. Let $\tilde{U}_{i,j}$ denote the jth entry of $\tilde{\mathbf{U}}_i$ and $\hat{\epsilon}_i(\mathbf{w})$ denote the ith entry of $\hat{\boldsymbol{\epsilon}}(\mathbf{w})$. We further define $\dot{\mathbf{G}}_{(s),i,j}\equiv\partial\mathbf{G}_{(s)}/\partial\tilde{U}_{i,j}$, $\ddot{\mathbf{G}}_{(s),i,j_1j_2}\equiv\partial^2\mathbf{G}_{(s)}/(\partial\tilde{U}_{i,j_1}\partial\tilde{U}_{i,j_2})$, $\dot{\mathbf{G}}_{i,j}(\mathbf{w})\equiv\Sigma_{s=1}^{S}w_s\dot{\mathbf{G}}_{(s),i,j}$ and $\ddot{\mathbf{G}}_{i,j_1j_2}(\mathbf{w})\equiv\Sigma_{s=1}^{S}w_s\ddot{\mathbf{G}}_{(s),i,j_1j_2}$. Straightforward but tedious calculation yields

$$\dot{\mathbf{G}}_{(s),i,j}=-\boldsymbol{\Pi}_{(s)}^{\mathrm{T}}\boldsymbol{\Lambda}_{(s)}\boldsymbol{\Upsilon}_{(s),i,j}\boldsymbol{\Lambda}_{(s)}\mathbf{Z}_{(s)}^{\mathrm{T}}+\boldsymbol{\Pi}_{(s)}^{\mathrm{T}}\boldsymbol{\Lambda}_{(s)}\boldsymbol{\Pi}_{(s)}\boldsymbol{\Sigma}^{1/2}\mathbf{h}_j\mathbf{b}_i^{\mathrm{T}} \tag{2}$$

and

$$
\begin{aligned}
\ddot{\mathbf{G}}_{(s),i,j_1 j_2} = {} & \boldsymbol{\Pi}_{(s)}^{\mathrm{T}} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Upsilon}_{(s),i,j_1} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Upsilon}_{(s),i,j_2} \boldsymbol{\Lambda}_{(s)} \mathbf{Z}_{(s)}^{\mathrm{T}} + \boldsymbol{\Pi}_{(s)}^{\mathrm{T}} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Upsilon}_{(s),i,j_2} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Upsilon}_{(s),i,j_1} \boldsymbol{\Lambda}_{(s)} \mathbf{Z}_{(s)}^{\mathrm{T}} \\
& - \boldsymbol{\Pi}_{(s)}^{\mathrm{T}} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Pi}_{(s)} \boldsymbol{\Sigma}^{1/2} (\mathbf{h}_{j_1} \mathbf{h}_{j_2}^{\mathrm{T}} + \mathbf{h}_{j_2} \mathbf{h}_{j_1}^{\mathrm{T}}) \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Pi}_{(s)}^{\mathrm{T}} \boldsymbol{\Lambda}_{(s)} \mathbf{Z}_{(s)}^{\mathrm{T}} \\
& - \boldsymbol{\Pi}_{(s)}^{\mathrm{T}} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Upsilon}_{(s),i,j_2} \boldsymbol{\Lambda}_{(s)} \boldsymbol{\Pi}_{(s)} \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_1} \mathbf{b}_i^{\mathrm{T}},
\end{aligned} \qquad (3)
$$

where $\boldsymbol{\Lambda}_{(s)} = (\mathbf{Z}_{(s)}^{\mathrm{T}} \mathbf{Z}_{(s)} - n\boldsymbol{\Sigma}_{(s)})^{-1}$ and $\boldsymbol{\Upsilon}_{(s),i,j} = \mathbf{Z}_{(s)}^{\mathrm{T}} \mathbf{b}_i \mathbf{h}_j^{\mathrm{T}} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Pi}_{(s)}^{\mathrm{T}} + \boldsymbol{\Pi}_{(s)} \boldsymbol{\Sigma}^{1/2} \mathbf{h}_j \mathbf{b}_i^{\mathrm{T}} \mathbf{Z}_{(s)}$. Now we can define

$$
C(\mathbf{w}) \equiv \|\hat{\epsilon}(\mathbf{w})\|^2 + 2\sigma^2 \operatorname{tr}\{\mathbf{Z}\mathbf{G}(\mathbf{w})\} - n\mathbf{Y}^{\mathrm{T}} \mathbf{G}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma} \mathbf{G}(\mathbf{w}) \mathbf{Y} + \sum_{l=1}^{5} A_l(\mathbf{w}), \qquad (4)
$$

where

$$
A_1(\mathbf{w}) = 2 \sum_{i=1}^{n} \sum_{j_1, j_2} \{\mathbf{Y}^{\mathrm{T}} \ddot{\mathbf{G}}_{i,j_1 j_2}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_1} \mathbf{Y}^{\mathrm{T}} \mathbf{G}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_2}\},
$$

$$
A_2(\mathbf{w}) = \sum_{i=1}^{n} \sum_{j_1, j_2} \{\mathbf{Y}^{\mathrm{T}} \dot{\mathbf{G}}_{i,j_2}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_1} \mathbf{Y}^{\mathrm{T}} \dot{\mathbf{G}}_{i,\, j_1}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_2} + \mathbf{Y}^{\mathrm{T}} \dot{\mathbf{G}}_{i,j_1}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_1} \mathbf{Y}^{\mathrm{T}} \dot{\mathbf{G}}_{i,j_2}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_{j_2}\},
$$

$$
A_3(\mathbf{w}) = -2 \sum_{i=1}^{n} \sum_{j=1}^{p} \{\hat{\epsilon}_i(\mathbf{w}) \mathbf{Y}^{\mathrm{T}} \dot{\mathbf{G}}_{i,j}(\mathbf{w})^{\mathrm{T}} \boldsymbol{\Sigma}^{1/2} \mathbf{h}_j\},
$$

$$
A_4(\mathbf{w}) = -2 \sum_{i=1}^{n} \sum_{j=1}^{p} \{\mathbf{Y}^{\mathrm{T}} \dot{\mathbf{G}}_{i,j}^{\mathrm{T}}(\mathbf{w}) \mathbf{Z}^{\mathrm{T}} \mathbf{b}_i \mathbf{Y}^{\mathrm{T}} \mathbf{G}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_j\},
$$

$$
A_5(\mathbf{w}) = -2\sigma^2 \sum_{i=1}^{n} \sum_{j=1}^{p} \{\mathbf{b}_i^{\mathrm{T}} \dot{\mathbf{G}}_{i,j}^{\mathrm{T}}(\mathbf{w}) \boldsymbol{\Sigma}^{1/2} \mathbf{h}_j\}.
$$

Although the definition of $C(\mathbf{w})$ appears complex, the idea behind it is actually quite simple. For selecting good weights, we need to compute the risk $R(\mathbf{w})$ as a function of $\mathbf{w}$. Intuitively, because $R(\mathbf{w})$ involves only the moments of various random variables, it should be able to be expressed explicitly in terms of the observations for a linear measurement error model. Thus, the focal point is in re-expressing $R(\mathbf{w})$, as is illustrated in the proof of theorem 1 given in section S.1.1 of the on-line supplementary material.

Theorem 1. For any weight $\mathbf{w}$, the criterion $C(\mathbf{w})$ is an unbiased estimator of the risk $R(\mathbf{w})$ up to $n\sigma^2$. Specifically

$$
R(\mathbf{w}) = E\{C(\mathbf{w})\} - n\sigma^2. \qquad (5)
$$

Theorem 1 indicates that, for selection of $\mathbf{w}$, we can ignore the offset $n\sigma^2$, which does not involve $\mathbf{w}$, and use $C(\mathbf{w})$ as if it were $R(\mathbf{w})$. For this, we shall minimize $C(\mathbf{w})$ with respect to $\mathbf{w}$ to select the optimal weights. Of course, $C(\mathbf{w})$ still involves the measurement error variance matrix $\boldsymbol{\Sigma}$ and the regression error variance σ^2. Thus, to implement the procedure in practice, we first need to obtain the estimates $\hat{\boldsymbol{\Sigma}}$ and $\hat{\sigma}^2$.

When σ^2 and $\boldsymbol{\Sigma}$ are both unknown, model (1) is not identifiable (Carroll *et al.*, 2006). Thus, to identify the model, additional information is always needed. In the measurement error literature, two main strategies are used to achieve identifiability: one is through using duplicate measurements corresponding to each $\mathbf{X}_i$, and the other is through introducing instrumental variables. Regardless of which strategy is implemented and what subsequent estimation procedure is used, the end product is a consistent estimator for $\boldsymbol{\Sigma}$, which is denoted as $\hat{\boldsymbol{\Sigma}}$. Thus, we base our following derivation on a variance–covariance estimator $\hat{\boldsymbol{\Sigma}}$, while omitting its detailed construction. We can extract the elements of $\hat{\boldsymbol{\Sigma}}$ to obtain estimates of $\boldsymbol{\Sigma}_{(s)}$ for $s \in \{1, \ldots, S\}$, denoted as $\hat{\boldsymbol{\Sigma}}_{(s)}$. Following Hansen (2007) and Wan *et al.* (2010), we estimate σ^2 on the basis of the model containing the largest number of covariates among the S candidate models. Assume that the index of the largest model is s^* and it contains p_{s^*} covariates. Then we estimate σ^2 by using $\hat{\sigma}^2 = \{\|\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(s^*)}\|^2 - n\hat{\boldsymbol{\beta}}_{(s^*)}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{(s^*)} \hat{\boldsymbol{\beta}}_{(s^*)}\}/(n - p_{s^*})$ (see page 155 of Carroll *et al.* (2006)). Plugging $\hat{\boldsymbol{\Sigma}}$ and $\hat{\sigma}^2$ into $C(\mathbf{w})$, a feasible weight choice criterion is

$$\hat{C}(\mathbf{w}) = C(\mathbf{w})|_{\sigma^2 = \hat{\sigma}^2, \boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}}. \tag{6}$$

We set the weights by minimizing $\hat{C}(\mathbf{w})$ with respect to $\mathbf{w}$ subject to $\Sigma_{s=1}^{S} w_s = 1$ and $w_s \geqslant 0$ for $i = 1, \ldots, S$, i.e.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \hat{C}(\mathbf{w}).$$

Remark 1. In our development of the weight choice criterion, we first assume that $\boldsymbol{\Sigma}$ is known and we introduce $C(\mathbf{w})$, and then we plug $\hat{\boldsymbol{\Sigma}}$ into $C(\mathbf{w})$ to form $\hat{C}(\mathbf{w})$. An alternative approach is to plug $\hat{\boldsymbol{\Sigma}}$ into $\hat{\boldsymbol{\beta}}$ first and then to form a new $R(\mathbf{w})$. One could then develop an unbiased estimator of the new $R(\mathbf{w})$ by using similar techniques to those in the proof of theorem 1, since $\hat{\boldsymbol{\Sigma}}$ generally depends on $\mathbf{Z}$. However, this alternative unbiased estimator will still depend on the unknown $\boldsymbol{\Sigma}$, while being more complicated than $C(\mathbf{w})$. In comparison, our current method of constructing the weight choice criterion bypasses this difficulty and is much simpler. In addition, as we shall show in theorem 2, our approach will yield optimal weight choice.

Remark 2. The unbiasedness result that is shown in theorem 1 relies heavily on the normality assumption of the measurement error $\mathbf{U}_i$. However, the optimality that is shown in theorem 2 and the consistency that is shown in theorem 3 do not need the normality assumption. In the simulation examples in Section 3, we find that, for non-normal measure error situations, our method also outperforms its competitors.

Remark 3. If we ignore the measurement errors, then $\mathbf{Z}_{(s)} = \mathbf{X}_{(s)}$ and $\mathbf{U}_{(s)} = \mathbf{0}$, by which we can take $\boldsymbol{\Sigma}_{(s)} = 0$ for $s \in \{1, \ldots, S\}$. Hence, by the definition of $C(\mathbf{w})$ in equation (6), we have

$$\hat{C}(\mathbf{w}) = \left\| \sum_{s=1}^{S} w_s \mathbf{Z}_{(s)}^{\mathrm{T}} (\mathbf{Z}_{(s)}^{\mathrm{T}} \mathbf{Z}_{(s)})^{-1} \mathbf{Z}_{(s)}^{\mathrm{T}} \mathbf{Y} - \mathbf{Y} \right\|^2 + 2\hat{\sigma}^2 (p_1, \ldots, p_S) \mathbf{w},$$

which is the Mallows criterion that was proposed by Hansen (2007).

It is easily seen that the criterion $\hat{C}(\mathbf{w})$ can be rewritten as $\hat{C}(\mathbf{w}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\Psi} \mathbf{w} + \mathbf{w}^{\mathrm{T}} \psi$ where $\boldsymbol{\Psi}$ is an $S \times S$ matrix and ψ is an S-dimensional vector. To minimize the quadratic function $\hat{C}(\mathbf{w})$ with respect to $\mathbf{w}$, there are many computational routines from various software packages. For example, in the R language it is solved by using the `quadprog` package, in MATLAB by the `quadprog` command and in SAS by the `qp` command. In our experience, they generally work effectively and efficiently even when S is very large. The computer code for our method is available from

https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets

2.3. Asymptotic optimality

In the linear regression framework without measurement error, it is known that minimizing the risk $R(\mathbf{w})$ leads to asymptotically optimal weights (Hansen, 2007). Considering the relationship between $C(\mathbf{w})$ and $R(\mathbf{w})$ in theorem 1, it is not surprising that minimizing $C(\mathbf{w})$ will lead to the same optimality property of the weights. Of course, because of the additional complexity that is caused by the measurement error as well as the need to approximate σ^2 and $\mathbf{\Sigma}$, it is much more difficult to establish such results. It also requires different conditions from the error-free case, as we now state.

Similarly to the definitions of $\mathbf{P}_{(s)}$, $\mathbf{P}(\mathbf{w})$, $L(\mathbf{w})$ and $R(\mathbf{w})$ defined before, we define these quantities in the error-free case. Specifically, let $\tilde{\mathbf{P}}_{(s)} = \mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}$, $\tilde{\mathbf{P}}(\mathbf{w}) = \Sigma_{s=1}^{S} w_s \tilde{\mathbf{P}}_{(s)}$, $\tilde{L}(\mathbf{w}) = \|\tilde{\mathbf{P}}(\mathbf{w})\mathbf{Y} - \boldsymbol{\mu}_0\|^2$ and $\tilde{R}(\mathbf{w}) = E\{\tilde{L}(\mathbf{w})\}$. In $\mathbf{P}_{(s)}$, $\mathbf{P}(\mathbf{w})$ and $L(\mathbf{w})$, we have replaced $\mathbf{\Sigma}$ by $\hat{\mathbf{\Sigma}}$, but for simplicity we still use this notation. Let $\lambda_{\max}(\mathbf{A})$ denote the maximum singular value for a matrix $\mathbf{A}$. We list the regularity conditions that are required for the asymptotic optimality of the weights chosen as stated above, where all the limiting properties here and throughout the text hold under $n \to \infty$.

Condition 1. $\mathbf{X}^{\mathrm{T}}\mathbf{X} = O(n)$, $\|\boldsymbol{\mu}_0\|^2 = O(n)$, $\lambda_{\max}(\mathbf{\Sigma}) < \infty$ and $E(\epsilon_i^4) < \infty$.

Condition 2. $\inf_{\mathbf{w}\in\mathcal{W}} \tilde{R}(\mathbf{w}) \to \infty$.

Condition 3. $n^{1/2}\sup_{\mathbf{w}\in\mathcal{W}}[\|\{\mathbf{P}(\mathbf{w}) - \tilde{\mathbf{P}}(\mathbf{w})\}\mathbf{Y}\|\tilde{R}^{-1}(\mathbf{w})] = o_p(1)$.

Condition 4. $\sup_{\mathbf{w}\in\mathcal{W}}[\|\mathbf{U}^{\mathrm{T}}\{\mathbf{P}(\mathbf{w}) - \mathbf{I}_n\}\mathbf{Y}\|^2\tilde{R}^{-2}(\mathbf{w})] = o_p(1)$.

Condition 5. $\sup_{\mathbf{w}\in\mathcal{W}}\{\lambda_{\max}(\mathbf{U}^{\mathrm{T}}\mathbf{U} - n\hat{\mathbf{\Sigma}})\tilde{R}^{-1}(\mathbf{w})\} = o_p(1)$.

Condition 1 is a standard condition for linear measurement error models, in which the restriction on the moments of ϵ requires the regression error distribution to have sufficiently thin tails. For example, it excludes the Cauchy distribution or Student t-distribution with degrees of freedom less than or equal to 4. Condition 2 is a general requirement that is necessary for the error-free linear regression model (Hansen, 2007; Liang *et al.*, 2011); hence it is also naturally imposed here. This condition is generally satisfied when none of the candidate models captures the true data generation procedure. Condition 3 requires the difference of $\mathbf{P}(\mathbf{w})$ and $\tilde{\mathbf{P}}(\mathbf{w})$ (both approximate a common quantity) to go to 0 uniformly relative to the risk in all different choices of weights. Similar conditions to condition 3 are used in other model averaging references, such as condition (A5) of Zhang *et al.* (2014). Condition 4 requires the covariance between the estimation residual and the measurement error to approach 0 relative to the risk in all different choices of weights. Finally, condition 5 requires the measurement error variance approximation to converge to the sample variance sufficiently fast in comparison with the risk. Conditions 4 and 5 are imposed so that the perturbations from the measurement error, once properly handled, do not overwhelm the signal in the risk calculation, which drives the model averaging process. It can be verified that, if $\hat{\mathbf{\Sigma}} - \mathbf{\Sigma} = O_p(n^{-1/2})$, the fourth moment of $\mathbf{U}_i$ exists and $n^{1/2}/\inf_{\mathbf{w}\in\mathcal{W}} \tilde{R}(\mathbf{w}) = o(1)$, then conditions 3–5 are implied by condition 1; the proof is in section S.1.3 of the on-line supplementary material.

Theorem 2 (asymptotic optimality). Under conditions 1–5,

$$\frac{L(\hat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w})} \to 1$$

in probability as $n \to \infty$.

Theorem 2 shows that the prescribed model averaging procedure is asymptotically optimal in the sense that its squared loss is asymptotically identical to that of the infeasible best possible model averaging estimator. The proof of theorem 2 is in section S.1.2 of the on-line supplementary material.

2.4. Consistency

Condition 2 generally excludes the situation that the true model is indeed linear. When none of the models being considered actually describes the data perfectly, it is natural that one seeks to average the imperfect candidate models to have performance that is superior to any single candidate model. However, there is also a possibility that the true model is indeed linear. In this case, it will be of interest to know what results from the model averaging procedure.

Assume that $\mu_{0i} = \mathbf{X}_i^{\mathrm{T}} \beta_0$, i.e. the true mean function μ_0 is indeed a linear function of the covariates with true parameter β_0. Here some or all elements of the true vector β_0 can be 0. The model averaging estimator of the regression parameter that is obtained from the method in Section 2.2 is naturally

$$\hat{\beta}(\hat{\mathbf{w}}) = \sum_{s=1}^{S} \hat{w}_s \mathbf{\Pi}_{(s)}^{\mathrm{T}} \hat{\beta}_{(s)}.$$

We now impose an additional condition concerning the measurement error structure. It is readily seen that it is a very mild condition and is easily satisfied except when the errors have very heavy tails.

Condition 6. $\hat{\mathbf{\Sigma}} - \mathbf{\Sigma} = O_p(n^{-1/2})$ and the fourth moment of $\mathbf{U}_i$ exists.

Theorem 3 (root n consistency). Under conditions 1 and 6, when $n \to \infty$,

$$\hat{\beta}(\hat{\mathbf{w}}) - \beta_0 = O_p(n^{-1/2}).$$

Theorem 3 complements the optimality property that was established in theorem 2. The two theorems reveal that the weight average modelling approach that we proposed here is optimal in terms of minimizing the relative loss when there does not exist a true regression parameter β_0, and it achieves root n convergence when there does exist a true parameter β_0. We, however, cannot establish the asymptotic distribution property of $\hat{\beta}(\hat{\mathbf{w}})$ or derive its asymptotic variance in the latter case because of the randomness of $\hat{\mathbf{w}}$. Much more research is needed in this area. The proof of theorem 3 is in section S.1.4 of the on-line supplementary material.

3. Simulation examples

3.1. Alternative methods

In this section, we conduct simulation experiments to demonstrate the finite sample performance of our model averaging method in linear measurement error models, MALMEM. We compare it with several other existing model averaging methods as well as several popular model selection methods. Two model selection methods in this context exist: AIC and BIC, which are widely used in the literature; see for example Liang and Li (2009) and Wang *et al.* (2012). Both methods select the model with the smallest criterion, defined as

$$C_{\mathrm{AIC}} = \|\mathbf{Y} - \mathbf{Z}_{(s)} \hat{\beta}_{(s)}\|^2 - n \hat{\beta}_{(s)}^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{(s)} \hat{\beta}_{(s)} + 2\hat{\sigma}^2 p_s$$

and

$$C_{\mathrm{BIC}} = \|\mathbf{Y} - \mathbf{Z}_{(s)} \hat{\beta}_{(s)}\|^2 - n \hat{\beta}_{(s)}^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{(s)} \hat{\beta}_{(s)} + \log(n) \hat{\sigma}^2 p_s.$$

The two existing model averaging methods were proposed in Buckland *et al.* (1997), where two weight choices were given, based respectively on the AIC and BIC mentioned above, and named the SAIC and SBIC. Specifically, the SAIC model average method assigns weights $w_{\mathrm{AIC},s} =$

$\exp(-C_{\mathrm{AIC}}/2)/\Sigma_{s=1}^{S}\exp(-C_{\mathrm{AIC}}/2)$ to model s and the SBIC model average method assigns weights $w_{\mathrm{BIC},s} = \exp(-C_{\mathrm{BIC}}/2)/\Sigma_{s=1}^{S}\exp(-C_{\mathrm{BIC}}/2)$ to model s.

3.2. Simulation designs

We consider two simulation settings. In the first, the true data generation procedure is captured by the candidate models, whereas, in the second, it is not. Hence, in the second setting, all candidate models are only approximations to the true data generation procedure.

3.2.1. Setting I

We generated data from model (1) with $\mu_{0i} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ and normal additive errors. Specifically, we set $n = 100, 200, 400$ and $p = 7$, and generated $\mathbf{X}_i = (x_{i,1}, \ldots, x_{i,7})^{\mathrm{T}}$ from a normal distribution with mean 0 and covariance $0.5^{|j_1 - j_2|}$ between x_{i,j_1} and x_{i,j_2}. We set $\boldsymbol{\Sigma} = \rho\mathbf{I}_p$, $\rho \in \{0.05, 0.2\}$, and $\boldsymbol{\beta}_0 = (1, 1, 0.5, 0, 0.3, -0.7, 0)^{\mathrm{T}}$ to generate $\mathbf{U}_i$, $\mathbf{Z}_i$ and Y_i. The parameter σ varies such that the theoretical $R^2 = \mathrm{var}(\mu_{0i})/\mathrm{var}(Y_i)$ varies in the set $\{0.1, 0.2, \ldots, 0.9\}$. We include two variables $x_{i,1}$ and $x_{i,2}$ in all candidate models. The five variables $x_{i,3}, \ldots, x_{i,7}$ are set to be auxiliary (i.e. they are possibly used in candidate models). This set-up is to mimic the situation that, in practice, some covariates are always set in candidate models based on theoretical or other grounds. Thus we have $2^5 = 32$ candidate models. To evaluate all five methods, we used 1000 replications and, in each replication, we computed model averaging estimators of $\boldsymbol{\mu}_0$ by $\hat{\boldsymbol{\mu}}(\hat{\mathbf{w}})$ and $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(\hat{\mathbf{w}}) = \Sigma_{s=1}^{S}\hat{w}_s\boldsymbol{\Pi}_{(s)}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(s)}$. Then, we computed risks as

$$\begin{aligned} L_\mu &= 1000^{-1}\sum_{r=1}^{1000}\|\hat{\boldsymbol{\mu}}(\hat{\mathbf{w}})^{(r)} - \boldsymbol{\mu}_0\|^2, \\ L_\beta &= 1000^{-1}\sum_{r=1}^{1000}\|\hat{\boldsymbol{\beta}}(\hat{\mathbf{w}})^{(r)} - \boldsymbol{\beta}_0\|^2, \end{aligned} \tag{7}$$

where $\hat{\boldsymbol{\mu}}(\hat{\mathbf{w}})^{(r)}$ and $\hat{\boldsymbol{\beta}}(\hat{\mathbf{w}})^{(r)}$ denote the estimator in the rth replication. To facilitate comparisons, all risks are normalized by the risk of the infeasible optimal estimator based on a single model. To check the performance of our method when measurement error is non-normal, we further set the distribution of $\mathbf{U}_i$ be uniform or χ^2; other setting are the same.

3.2.2. Setting II

This design is based on the setting of Hansen (2007), except that covariates are subject to measurement error. Specifically, we generated data from model (1) with $\mu_{0i} = \Sigma_{j=1}^{\infty}x_{ij}\beta_j$ and normal additive errors. We set $x_{i1} = 1$ and observations of all other x_{ij}s are generated from the $N(0, 1)$ distribution and are independent. The coefficients $\beta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 0.5$. The sample size varies as 100, 200 and 400. The number of approximating models is $S = 18$. The sth candidate model contains the first s observed covariates. We used $\boldsymbol{\Sigma} = \rho\mathbf{I}_{S-1}$ and $\rho \in \{0.05, 0.2\}$ to generate $\mathbf{U}_i$ and $\mathbf{Z}_i$. For the intercept x_{i1}, there is no measurement error. In this setting, following Hansen (2007), we compare the five methods based on their L_μ-values in expression (7). To address the comments of the referees that one may ignore measurement errors if the focus is on prediction, we also compare our method with Mallows model averaging, which was introduced in remark 3.

3.3. Simulation results

The results of the simulations are given in Figs 1 and 2 for setting I and in Fig. 3 for setting II. A summary of these results is very simple. In almost all cases, and generally, our method

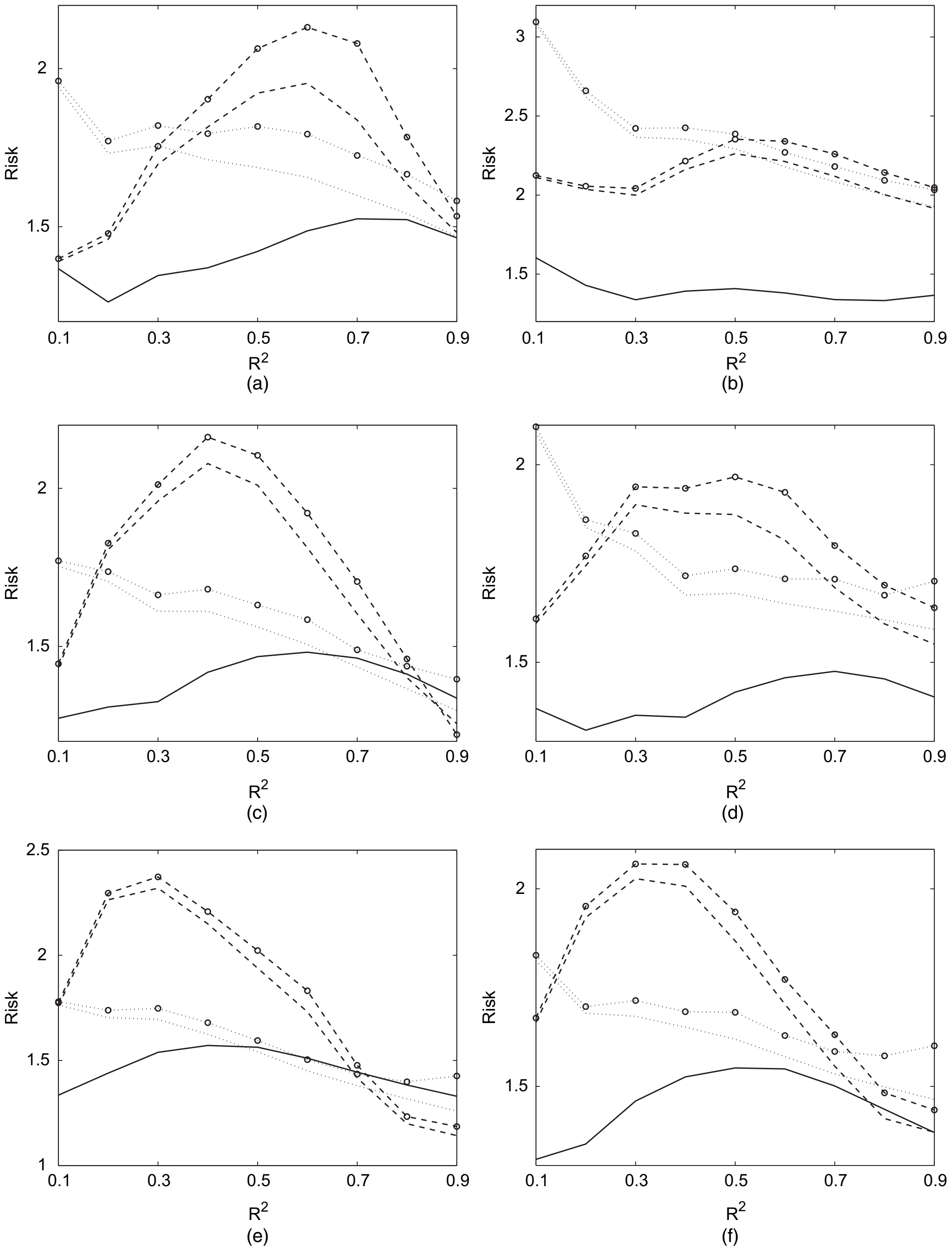
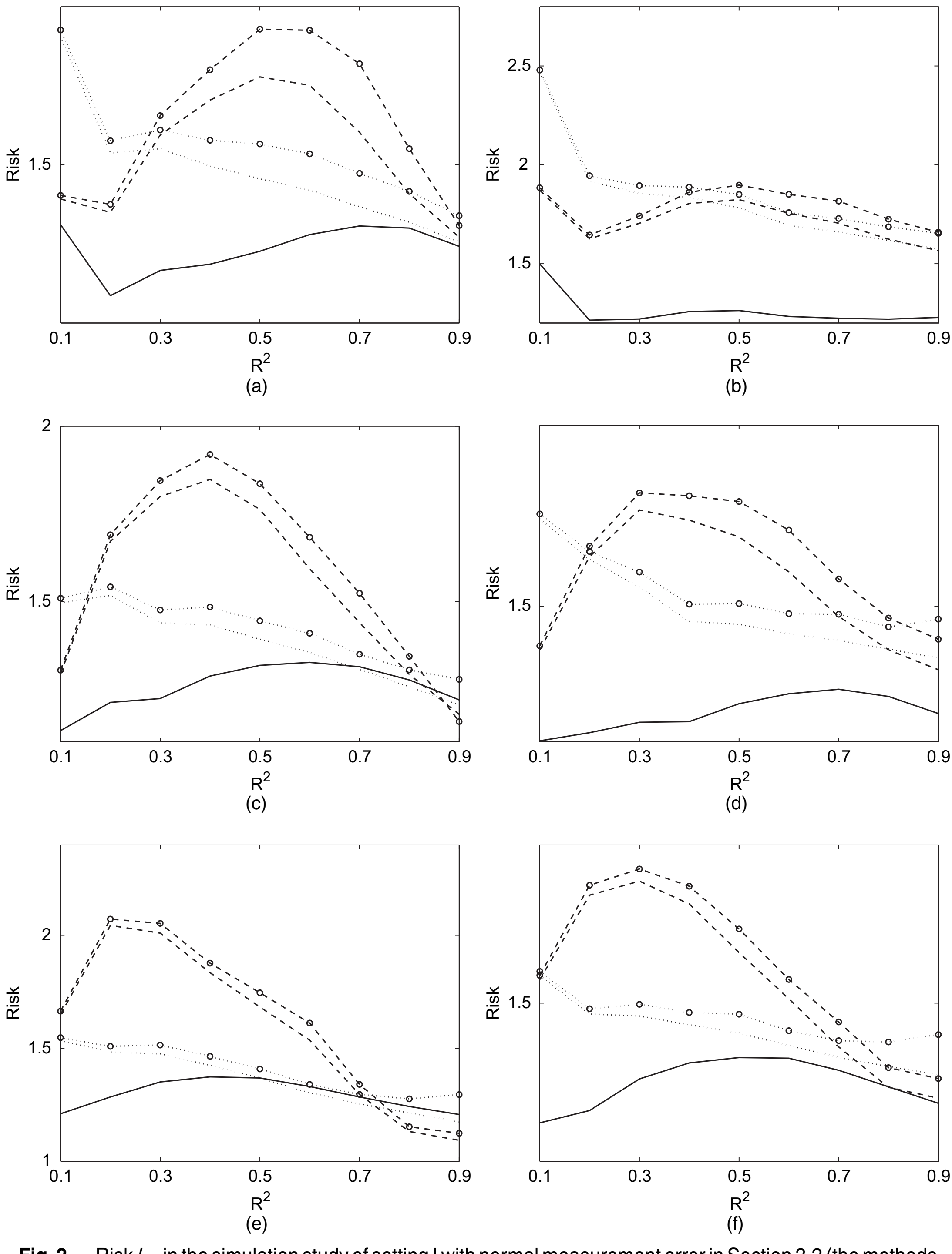


Fig. 1. Risk L_β in the simulation study of setting I with normal measurement error in Section 3.2 (the methods compared are AIC- (○) and BIC- (⊖) based model selection, SAIC- (••••••) and SBIC- (– – –) based model averaging and our asymptotically optimal model averaging method MALMEM (———): (a) $n=100$, $\rho=0.05$; (b) $n=100$, $\rho=0.2$; (c) $n=200$, $\rho=0.05$; (d) $n=200$, $\rho=0.2$; (e) $n=400$, $\rho=0.05$; (f) $n=400$, $\rho=0.2$



(a)

(b)

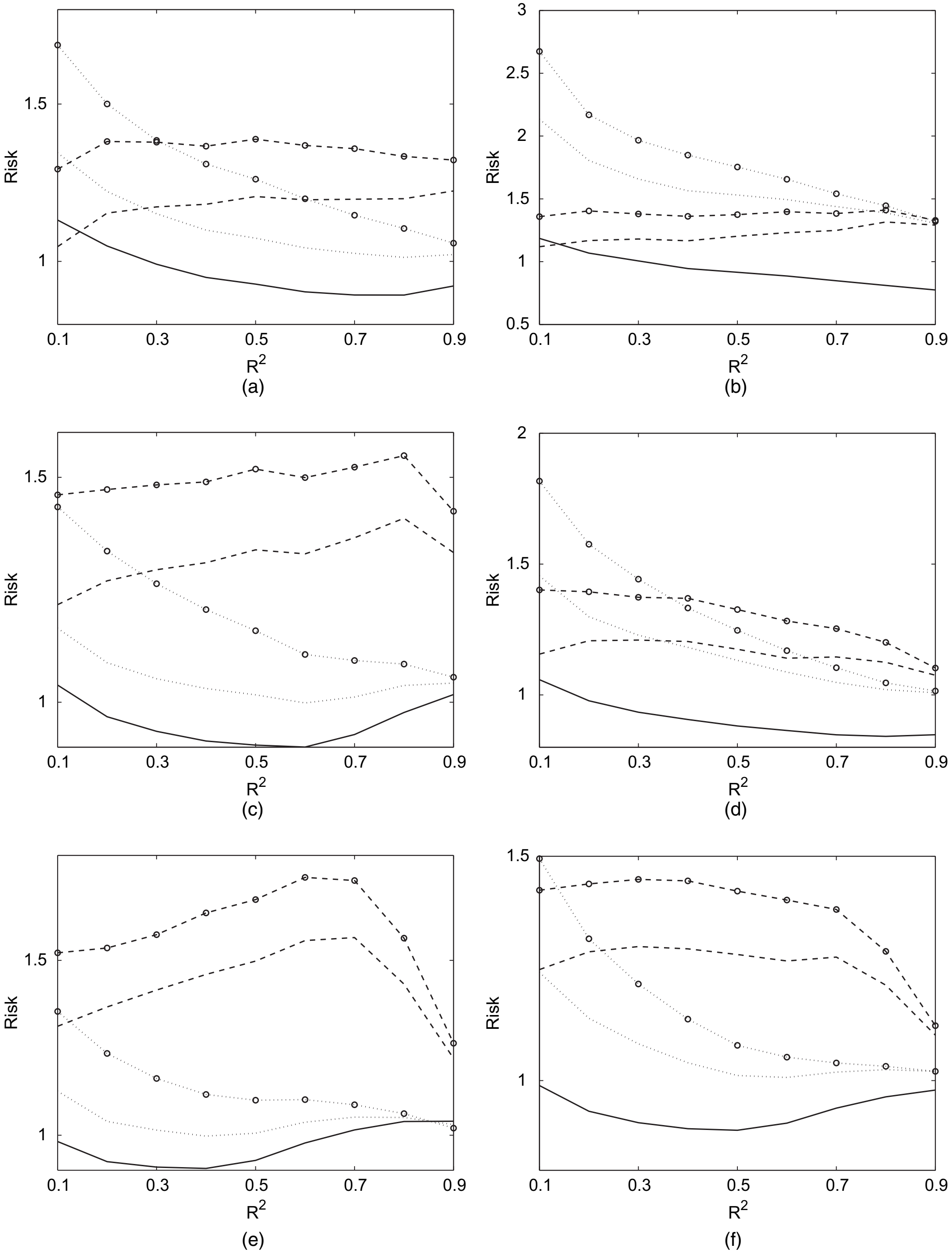
(c)

(d)

(e)

(f)

Fig. 2. Risk L_μ in the simulation study of setting I with normal measurement error in Section 3.2 (the methods compared are AIC- (○) and BIC- (⊖) based model selection, SAIC- (••••••) and SBIC- (– – –) based model averaging and our asymptotically optimal model averaging method MALMEM (———): (a) $n = 100$, $\rho = 0.05$; (b) $n = 100$, $\rho = 0.2$; (c) $n = 200$, $\rho = 0.05$; (d) $n = 200$, $\rho = 0.2$; (e) $n = 400$, $\rho = 0.05$; (f) $n = 400$, $\rho = 0.2$



(a) (b) (c) (d) (e) (f)

Fig. 3. Risk L_μ in the simulation study of setting II in Section 3.2 (the methods compared are AIC- (○) and BIC- (⊖) based model selection, SAIC- (•••••••) and SBIC- (– – –) based model averaging and our asymptotically optimal model averaging method MALMEM (———): (a) $n=100$, $\rho=0.05$; (b) $n=100$, $\rho=0.2$; (c) $n=200$, $\rho=0.05$; (d) $n=200$, $\rho=0.2$; (e) $n=400$, $\rho=0.05$; (f) $n=400$, $\rho=0.2$

MALMEM greatly dominates the other methods. When R^2 is very high, the selection methods by AIC and BIC can be better than model averaging methods. The possible reason is that the small noise in the data enables the selection criteria to choose the best model with very high frequencies. These numerical results are not unexpected given our theoretical results, although the magnitude of the improvement in risk was somewhat unexpected.

The numerical results with non-normal measurement errors under setting I are shown in Figs S.1–S.4 of the on-line supplementary material, from which we find that MALMEM still greatly dominates the other methods in most cases. Hence, the performance of our method is not sensitive to the distribution of measurement errors. The comparison results between MALMEM with the Mallows model averaging method are shown in Fig. S.5 of the supplementary material, which shows that MALMEM outperforms Mallows model averaging in most cases especially when R^2 is large. To check whether this performance is sensitive to the distribution of measurement error, we further set the distribution of $\mathbf{U}_i$ to be uniform or χ^2, $n=200$, and keep the other settings. Fig. S.6 of the supplementary material shows that the non-normal measurement errors change their performance very slightly.

4. Application to health study

4.1. The data

The theory in Section 2 and the simulations in Section 3 suggest that, when we apply our procedure MALMEM to actual data, we should see large gains in predictive accuracy: something which will be confirmed in Section 4.4.

We analyse data from the Women's Interview Study of Health (WISH) (Potischman *et al.*, 1999). The data that we use here consist of 1209 healthy women who did not develop breast cancer. Each woman completed a food frequency questionnaire, from which we collected the measurements of daily intakes of protein, Z_4, fat, Z_5, and carbohydrates, Z_6. In addition to the main study, a subset of these women participated in a validation study where, for 12 days, their dietary intakes were measured by a combination of 24-h recalls and dietary records: first six randomly selected days with 24-h recalls, and then two randomly selected 3-day periods of dietary records. We therefore use these 12 additional measurements to form the true protein, fat and carbohydrate intakes. We used the cube root of the dietary data, which in all cases is far more normally distributed than in the original scale.

We then standardized the variables to the first day of dietary records so that each day had the same mean and standard deviation. This was done as in equation (3) of Nusser *et al.* (1996), so that, if a_j and b_j are the sample mean and sample standard deviation on day j, and the individual measurements are V_{ij}, then $V_{ij}^* = a_7 + (b_7/b_j)(V_{ij} - a_j)$, where day 7 is the first dietary record. Then the resulting 12-day average $\Sigma_{j=1}^{p} V_{ij}^*/12$ is taken as the *true* intake X_i for each individual. Similar methodology has been used for obtaining such true intakes, e.g. Spiegelman *et al.* (2001) and Yi *et al.* (2015).

There are a total of 178 subjects in the validation study. From the true intakes in the validation study and the intakes with measurement error in the main study, we can obtain the measurement error variance–covariance matrix that is associated with $(Z_4, Z_5, Z_6)^{\mathrm{T}}$. We also included age, Z_1, and a discrete variable with three levels, from which we develop two dummy variables, Z_2 meaning smoking status 1 (past smoker) and Z_3 meaning smoking status 2 (current smoker). Except for the two dummy variables Z_2 and Z_3, we performed a transformation on all other variables by taking a cube root. The response variable that we consider here is the cube root of the body mass index BMI, the transformation making the response far less skewed. Thus, we have a total of six covariates, and we consider $2^6 - 1 = 63$ candidate models.

4.2. Comparison of models

Table 1 contains the AIC and BIC values and the model averaging weights of various candidate models. We list only the candidate models whose largest weights for all model averaging and selection methods are at least 0.01. Here, we used the indices of the covariates to indicate which variables are included in a candidate model. For example, (4, 5) indicates that the model includes protein Z_4 and fat Z_5. The methods AIC, BIC, SAIC and SBIC are the same as those defined in Section 3. We can see that the BIC selects model (2, 5, 6) and the BIC value of model (2, 3, 5, 6) is close to that of model (2, 5, 6). The AIC also supports models (2, 5, 6) and (2, 3, 5, 6), with their AIC values identical to the two digits provided. This indicates that the AIC and BIC cannot clearly identify a best model for the data. Similarly, the SAIC and SBIC both assign very similar weights to model (2, 5, 6) and model (2, 3, 5, 6). In contrast, MALMEM clearly favours model (2, 5). It assigns a large weight of 0.76 to model (2, 5), whereas it assigns a weight of 0.06 to model (4, 5, 6), 0.09 to model (2, 4, 5, 6) and 0.09 to model (3, 4, 5, 6).

We shall show later in Section 4.4 that the MALMEM solution leads to vastly better predictive accuracy in the validation data.

Table 1. Analysis of the WISH data of Section 4†

Model	*Model selection criterion values*		*Weights*		
	AIC	*BIC*	*SAIC*	*SBIC*	*MALMEM*
(2, 5)	−35.76	−34.94	0.00	0.00	*0.76*
(2, 5, 6)	*−79.20*	*−77.97*	*0.32*	*0.38*	0.00
(4, 5, 6)	−47.12	−45.89	0.00	0.00	0.06
(2, 3, 5, 6)	*−79.20*	*−77.56*	*0.32*	*0.31*	0.00
(2, 4, 5, 6)	−77.98	−76.34	0.18	0.17	0.09
(3, 4, 5, 6)	−50.14	−48.50	0.00	0.00	0.09
(2, 3, 4, 5, 6)	−78.01	−75.96	0.18	0.14	0.00

†AIC and BIC values and weights by SAIC- and SBIC-based model averaging and our method (MALMEM). The variables in order were age Z_1, past smoker Z_2, current smoker Z_3, protein Z_4, fat Z_5 and carbohydrates Z_6. Models such as (2, 5) mean that Z_2 and Z_5 were in the model.

Table 2. Analysis of the WISH data of Section 4†

Model (2, 5), weight = 0.76	Covariates	2	5		
	Estimate	−0.16	0.76		
	Standard deviation	0.02	< 0.005		
Model (4, 5, 6), weight = 0.06	Covariates	4	5	6	
	Estimate	−0.05	1.82	−0.75	
	Standard deviation	0.14	0.30	0.17	
Model (2, 4, 5, 6), weight = 0.09	Covariates	2	4	5	6
	Estimate	−0.44	−0.19	2.80	−1.35
	Standard deviation	0.06	0.15	0.43	0.25
Model (3, 4, 5, 6), weight = 0.09	Covariates	3	4	5	6
	Estimate	0.13	−0.05	1.90	−0.81
	Standard deviation	0.03	0.14	0.31	0.18

†Displayed are the parameter estimates and their standard deviations for the four models with weights larger than 0.01: the model weights are also displayed. The variables in order were age Z_1, past smoker Z_2, current smoker Z_3, protein Z_4, fat Z_5 and carbohydrates Z_6. Models such as (2, 5), the model assigned a weight of 0.76, mean that Z_2 and Z_5 were in the model.

4.3. The MALMEM averaged model and interpretation

To interpret the average model more clearly, we computed the effect sizes of each of the continuous variables, namely the model average fit when the true dietary variables were standardized each to have variance 1.0. This involves division of the observed data for protein, fat and carbohydrates by 0.3485, 0.4467 and 0.578 respectively. Call the standardized variables stProtein, stFat and stCarbohydrates. The final averaged model fit is

$$-0.16\, I(\text{past smoker}) + 0.01\, I(\text{current smoker}) - 0.01\, \text{stProtein} \\ + 0.50\, \text{stFat} - 0.14\, \text{stCarbohydrates}. \qquad (8)$$

We see from expression (8) and Table 1 that the effect size of fat intake is far larger than any of the other variables, and the positive sign of the coefficient is expected. For carbohydrates and BMI, Gaesser (2007) stated that their relationship is controversial, but that a

> 'review of relevant literature indicates that most epidemiologic studies show an inverse relationship between carbohydrate intake and BMI, even when controlling for potential confounders'.

Thus the negative sign for carbohydrates in expression (8) is supported by relevant literature.

4.4. Comparison of predictive performance

Finally, and importantly, we use the validation data to check the prediction performance of the model selection and averaging methods. The boxplots of squared prediction errors based on the five methods are displayed in Fig. 4, where we can see that the boxplot corresponding to MALMEM has by far the best performance: for example, the median and the 75th percentiles of the other methods are nearly eight times larger than that of MALMEM.

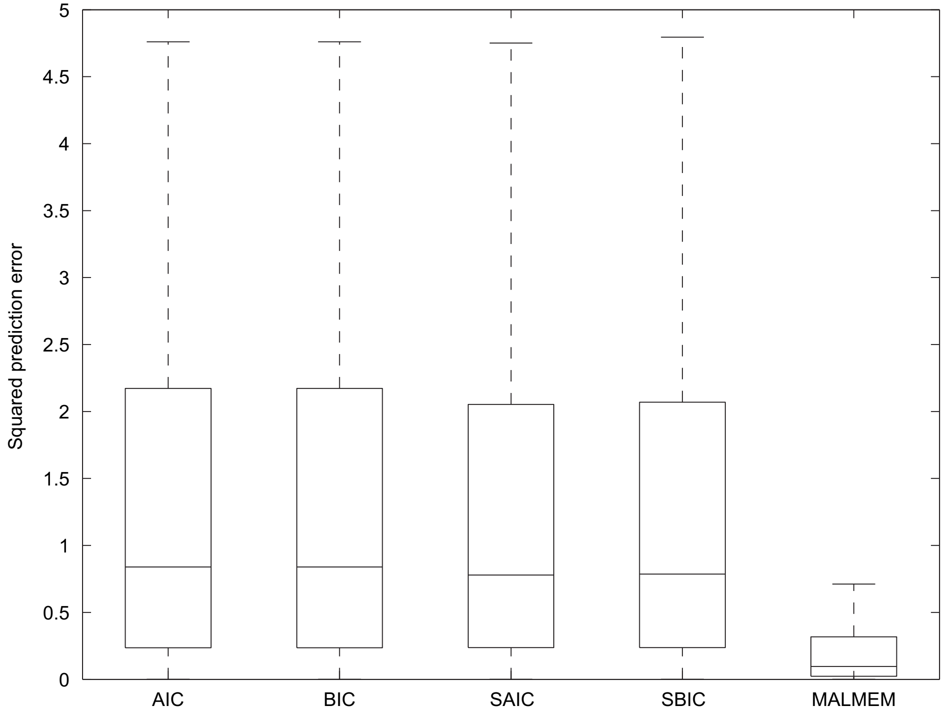


Fig. 4. Analysis of the WISH data: boxplots of 178 squared prediction errors (the methods compared are AIC- and BIC-based model selection, SAIC- and SBIC-based model averaging and our method MALMEM

5. Discussion

We have proposed a model averaging method, called MALMEM, for linear measurement error models. When the true model is not included in the set of candidate models, the method was shown to be asymptotical optimal in the sense of achieving the lowest squared loss in large sample sizes, whereas, when the true model is included, the parameter estimates by the model averaging method are root n consistent. Numerical analysis in comparison with existing model selection methods strongly favours MALMEM. MALMEM was applied to the WISH.

We have assumed that the dimension of candidate model p_s and the number of candidate models S are fixed when the sample size n increases. When p_s and S increase with n, the unbiasedness property in theorem 1 still holds. However, more restrictive conditions will be needed for asymptotic optimality: the problem certainly needs more careful further investigation.

As for all model averaging methods, if the number of covariates is large, possible candidate models are numerous; hence the computation of the procedure will be cumbersome. In this case, a model screening step before model averaging is desirable. The AIC and BIC have been used in screening steps of Yuan and Yang (2005) and Zhang *et al.* (2013), and Claeskens *et al.* (2006) and Zhang *et al.* (2012) suggested the use of stepwise screening.

We use additivity of the measurement error in developing the weight choice criterion. When the measurement error is multiplicative, one choice is transforming it to be additive by taking logarithms. Another choice is directly developing a model averaging method for the situation with multiplicative measurement errors, or even more general error structures. This will be very different from the method that is developed in the current paper because the coefficient estimators will be very different (see Hwang (1986)) and warrants further investigation.

Acknowledgements

Zhang's research was supported by the National Natural Science Foundation of China (grants 71522004, 11471324 and 71631008). Ma's research was partially supported by grants from the National Science Foundation and National Institutes of Health. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030). The work of Zhang was performed during his visits to Texas A&M University and Pennsylvania State University.

References

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997) Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.

Buonaccorsi, J. P. (2010) *Measurement Error: Models, Methods and Applications*. New York: Chapman and Hall.

Carroll, R. J., Delaigle, A. and Hall, P. (2009) Nonparametric prediction in measurement error models. *J. Am. Statist. Ass.*, **104**, 993–1014.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006) *Measurement Error in Nonlinear Models: a Modern Perspective*. Boca Raton: Chapman and Hall–CRC.

Chen, J., Li, D., Linton, O. and Lu, Z. (2018) Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *J. Am. Statist. Ass.*, **113**, 919–932.

Cheng, X. and Hansen, B. E. (2015) Forecasting with factor-augmented regression: a frequentist model averaging approach. *J. Econmetr.*, **186**, 280–293.

Claeskens, G., Croux, C. and van Kerckhoven, J. (2006) Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, **62**, 972–979.

Fang, F., Lan, W., Tong, J. and Shao, J. (2019) Model averaging for prediction with fragmentary data. *J. Bus. Econ. Statist.*, **37**, 1–11.

Fuller, W. (1987) *Measurement Error Models*. New York: Wiley.

Gaesser, G. A. (2007) Carbohydrate quantity and quality in relation to body mass index. *J. Am. Dietet. Ass.*, **107**, 1768–1780.

Gustafson, P. (2004) *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton: Chapman and Hall–CRC.

Hansen, B. E. (2007) Least squares model averaging. *Econometrica*, **75**, 1175–1189.

Hansen, B. E. (2009) Averaging estimators for regressions with a possible structural break. *Econmetr. Theory*, **25**, 1498–1514.

Hansen, B. E. and Racine, J. (2012) Jacknife model averaging. *J. Econmetr.*, **167**, 38–46.

Hjort, N. L. and Claeskens, G. (2003) Frequentist model average estimators. *J. Am. Statist. Ass.*, **98**, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**, 382–417.

Hwang, J. T. (1986) Multiplicative errors-in-variables models with applications to recent data released by the US Department of Energy. *J. Am. Statist. Ass.*, **81**, 680–688.

Li, D., Linton, O. and Lu, Z. (2015) A flexible semiparametric forecasting model for time series. *J. Econmetr.*, **187**, 345–357.

Liang, H. and Li, R. (2009) Variable selection for partially linear models with measurement errors. *J. Am. Statist. Ass.*, **104**, 234–248.

Liang, H., Zou, G., Wan, A. T. K. and Zhang, X. (2011) Optimal weight choice for frequentist model average estimators. *J. Am. Statist. Ass.*, **106**, 1053–1066.

Liu, C.-A. (2015) Distribution theory of the least squares averaging estimator. *J. Econmetr.*, **186**, 42–159.

Liu, Q. and Okui, R. (2013) Heteroskedasticity-robust Cp model averaging. *Econmetr. J.*, **16**, 462–473.

Longford, N. T. (2005) Model selection and efficiency—is 'Which model ...?' the right question? *J. R. Statist. Soc.* A, **168**, 469–472.

Lu, X. and Su, L. (2015) Jackknife model averaging for quantile regressions. *J. Econmetr.*, **188**, 40–58.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W. and Fuller, W. A. (1996) A semiparametric transformation approach to estimating usual daily intake distributions. *J. Am. Statist. Ass.*, **91**, 1440–1449.

Potischman, N., Carroll, R. J., Iturria, S. J., Mittl, B., Curtin, J., Thompson, F. E. and Brinton, L. A. (1999) Comparison of the 60- and 100-item NCI-block questionnaires with validation data. *Nutrn Cancer*, **34**, 70–75.

Spiegelman, D., Carroll, R. J. and Kipnis, V. (2001) Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statist. Med.*, **20**, 139–160.

Wan, A. T. K., Zhang, X. and Zou, G. (2010) Least squares model averaging by Mallows criterion. *J. Econmetr.*, **156**, 277–283.

Wang, H., Zou, G. and Wan, A. T. K. (2012) Model averaging for varying-coefficient partially linear measurement error models. *Electron. J. Statist.*, **6**, 1017–1039.

Yang, Y. (2001) Adaptive regression by mixing. *J. Am. Statist. Ass.*, **96**, 574–588.

Yi, G., Ma, Y., Spiegelman, D. and Carroll, R. J. (2015) Functional and structural methods with mixed measurement error and misclassification in covariates. *J. Am. Statist. Ass.*, **110**, 681–696.

Yuan, Z. and Yang, Y. (2005) Combining linear regression models: when and how? *J. Am. Statist. Ass.*, **100**, 1202–1214.

Zhang, X. (2013) Model averaging with covariates that are missing completely at random. *Econ. Lett.*, **121**, 360–363.

Zhang, X. and Liang, H. (2011) Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.*, **39**, 174–200.

Zhang, X., Lu, Z. and Zou, G. (2013) Adaptively combined forecasting for discrete response time series. *J. Econmetr.*, **176**, 80–91.

Zhang, X., Wan, A. T. K. and Zhou, S. Z. (2012) Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *J. Bus. Econ. Statist.*, **30**, 132–142.

Zhang, X., Yu, D., Zou, G. and Liang, H. (2016) Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *J. Am. Statist. Ass.*, **111**, 1775–1790.

Zhang, X., Zou, G. and Liang, H. (2014) Model averaging and weight choice in linear mixed-effects models. *Biometrika*, **101**, 205–218.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for "MALMEM: model averaging in linear measurement error models"'.