Model Averaging for Varying-Coefficient Partially Linear Measurement Error Models

Haiying Wang[1,2], Guohua Zou[2], and Alan T. K. Wan[3]

May 25, 2012

[1]*Department of Statistics, University of Missouri, Columbia, Missouri 65211, USA*
[2]*MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, P. R. China*
[3]*Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong*

Abstract

In a 2003 paper, Hjort and Claeskens proposed a framework for studying the limiting distributions and asymptotic risk properties of model average estimators under parametric models. They also suggested a simple method for constructing confidence intervals for the parameters of interest estimated by model averaging. The purpose of this paper is to broaden the scope of the aforementioned study to include a semi-parametric varying-coefficient partially linear measurement error model. Within this context, we develop a model averaging scheme for the unknowns, derive the model average estimator's asymptotic distribution, and develop a confidence interval procedure of the unknowns with an actual coverage probability that tends toward the nominal level in large samples. We further show that confidence intervals that are constructed based on the model average estimators are asymptotically the same as those obtained under the full model. A simulation study examines the finite sample performance of the model average estimators, and a real data analysis illustrates the application of the method in practice.

Keywords and phrases: asymptotic equivalence, measurement errors, model averaging, model selection, semi-parametric models

1 Introduction

Model selection has always been an integral part of statistical analysis. Well-known criteria for model selection include the AIC (Akaike, 1973), Mallows' C_p (Mallows, 1973), Cross Validation (Stone, 1974), BIC (Schwarz, 1978), Generalized Cross Validation (Craven and Wahba, 1979), RIC (Foster and George, 1994), FIC (Claeskens and Hjort, 2003), among others. The search for the "best" model recognizes the existence of more than one plausible model structure, implying a level of uncertainty associated with the choice of model. However, this uncertainty is usually ignored when it comes to making an inference contingent on

the chosen best model, and this results in an overconfident inference about the unknowns (Hjort and Claeskens, 2003; Danilov and Magnus, 2004; Claeskens and Hjort, 2008; Liang *et al.*, 2011). It is also well-known that many model selection techniques can be highly influenced by slight variations in data.

One method for incorporating model uncertainty in statistical analysis is model averaging; instead of selecting a single model, model averaging compromises across the set of plausible models, weighted by some criteria that reflect the degree to which each model is trusted. Bayesian model averaging (BMA) has been promoted in a range of disciplines as a means of incorporating model uncertainty. Excellent surveys of the vast BMA literature can be found in Draper (1995), Raftery *et al.* (1997), Hoeting *et al.* (1999), and Clyde and George (2004). A key component of BMA is the use of prior distributions of the unknowns and models. While this provides a formal framework for incorporating prior knowledge of the process being modeled, any poor handling of prior distributions can lead to undesirable behavior of the posterior distributions and model average estimator. Frequentist model averaging (FMA), on the other hand, precludes the need to specify any prior distribution, although how to determine an optimal weight choice by a data-driven approach is arguably the biggest challenge for the frequentist formulation.

Compared to the immense amount of BMA literature, the literature on FMA is more recent, nonetheless a great deal of work has been invested in developing model weighting schemes for FMA estimators and the investigation of their properties. The early work of Buckland *et al.* (1997) described an approach that uses the exponent of the negative of the AIC value as the weight for an individual model. Yang (2003) and Yuan and Yang (2005) developed an adaptive regression by mixing (ARM) algorithm, while Leung and Barron (2006) proposed a weight choice criterion based on risk minimization. More recently, Hansen (2007, 2008) and Wan *et al.* (2010) developed an FMA based on the Mallows' criterion. Of particular relevance to the current study is the work of Hjort and Claeskens (2003), who developed an asymptotic theory for frequentist model averaging in parametric models based on a local mis-specification framework, which shows that FMA generally results in an estimator with a non-normal asymptotic distribution. They also suggested a simple method for confidence interval construction of the unknown parameters. Hjort and Claeskens' (2003) analysis has been extended to several other models including the Cox's hazard model (Hjort and Claeskens, 2006), general semi-parametric models (Claeskens and Carroll, 2007), the generalized additive partial linear model advanced by Zhang and Liang (2011), and the censored regression model (Zhang *et al.*, 2012). A summary of these recent developments can be found in Wang *et al.* (2009).

The current paper extends Hjort and Claeskens' (2003) investigation to the varying-coefficient partially linear measurement error (VCPLE) model. The varying-coefficient partially linear (VCPL) model (Zhang *et al.*, 2002; Fan and Huang, 2005) allows the different covariates in the model to interact in a flexible way and has been an important development in the semi-parametric literature in recent years. It also covers many other semi-parametric models including the varying-coefficient model (Hastie and Tibshirani, 1993) and the partially linear model (Engle *et al.*, 1986) as special cases. The VCPLE model considered in this paper is a version of the VCPL model where the covariates in the parametric component of the model are measured with additive errors. The VCPLE model was previously considered by You and Chen (2006), who suggested an alternative estimation procedure that leads to

consistent estimators of the parametric and non-parametric components of the model. In this paper we are concerned with model averaging within the VCPLE framework; in particular, we focus on the derivation of the model average estimator's asymptotic distribution, and develop a method for constructing confidence intervals of the unknowns along the lines of Hjort and Claeskens (2003). We demonstrate that these confidence intervals have a coverage probability that tends toward the nominal level in large samples. We also prove that the FMA-based confidence intervals are asymptotically the same as the confidence intervals based on the full model.

The remainder of the paper is organized as follows. Section 2 presents the model setup and discusses the estimation method of the unknowns in each candidate model. Section 3 describes the model averaging scheme and presents the main theoretical results. Section 4 reports the results of a simulation study that examines the finite sample performance of the model average estimator. Section 5 applies the proposed method to a real data set on dietary intake measurements. Section 6 presents the conclusion. The appendix contains the proofs of lemmas and theorems.

2 Model setup and estimation methods

Consider the i.i.d. samples (Y_i, W_i, Z_i, T_i), $i = 1, ..., n$, and the following VCPLE model:

$$\begin{cases} Y_i &= X_i^\top \theta + Z_i^\top \alpha(T_i) + \varepsilon_i, \\ W_i &= X_i + U_i, \end{cases} \tag{1}$$

where Y_i is the response variable, (X_i, Z_i, T_i) are covariates, $\theta = (\beta^\top, \gamma^\top)^\top$ with β and γ being p and q dimensional coefficient vectors respectively, $\alpha(\cdot) = \{\alpha_1(\cdot), ..., \alpha_r(\cdot)\}^\top$ is an r dimensional unknown coefficient function, and ε_i is a random error with mean 0 and variance σ^2, and independent of (X_i, Z_i, T_i). As in Fan and Huang (2005) and You and Chen (2006), we assume that the dimension of T_i is one. Here, it is assumed that X_i cannot be observed, but instead its surrogate W_i is observed, with U_i being a vector of random errors with mean 0 and covariance matrix Σ_u; further, U_i is independent of (X_i, Z_i, T_i) and ε_i. For analytical convenience we assume throughout our theoretical analysis that Σ_u is known. This last assumption does not give rise to any loss of generality because all results continue to hold if Σ_u is replaced by a consistent estimator when Σ_u is unknown.

Clearly, when $U_i \equiv 0$, the VCPLE model reduces to the VCPL model which contains many common models as special cases. For example, when $\theta \equiv 0$, it reduces to the varying-coefficient model; when $r = 1$ and $Z_i \equiv 1$, it becomes the partially linear model. Write $\mathbf{Y} = (Y_1, ..., Y_n)^\top$, $\mathbf{X} = (X_1, ..., X_n)^\top$, $\mathbf{W} = (W_1, ..., W_n)^\top$, $\mathbf{U} = (U_1, ..., U_n)^\top$, $\mathbf{Z} = (Z_1, ..., Z_n)^\top$, $\mathbf{T} = (T_1, ..., T_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$ and $\mathbf{M} = \{Z_1^\top \alpha(T_1), ..., Z_n^\top \alpha(T_n)\}^\top$, then model (1) can be expressed as

$$\begin{cases} \mathbf{Y} = \mathbf{X}\theta + \mathbf{M} + \boldsymbol{\varepsilon}, \\ \mathbf{W} = \mathbf{X} + \mathbf{U}. \end{cases} \tag{2}$$

When there are no measurement errors, the profile least-squares method described in Fan and Huang (2005) can be used to estimate θ. To estimate $\alpha_j(t)$, write, for any given

θ, $Y_i^* = Y_i - X_i^\top \theta$. Then model (1) becomes the general varying-coefficient model, and the following local linear approximation can be used to estimate $\alpha_j(t)$:

$$\alpha_j(t_0) + \alpha_j'(t_0)(t - t_0) \equiv a_j + b_j(t - t_0),\ j = 1, 2, ..., r,$$

for any t in the neighborhood of t_0.

Denote $\mathbf{a} = (a_1, ..., a_r)^\top$ and $\mathbf{b} = (b_1, ..., b_r)^\top$. Then $\mathbf{a}$ and $\mathbf{b}$ can be estimated by the local weighted least-squares method based on the criterion

$$\min_{\mathbf{a},\mathbf{b}} \sum_{i=1}^{n} \left[Y_i^* - Z_i^\top \left\{\mathbf{a} + \mathbf{b}(T_i - t_0)\right\}\right]^2 K_h(T_i - t_0),$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is a bandwidth. Write the solution to this minimization problem as

$$\left\{\hat{a}_1(t), ..., \hat{a}_r(t), h\hat{b}_1(t), ..., h\hat{b}_r(t)\right\}^\top = (\mathcal{D}_t^\top \Omega_t \mathcal{D}_t)^{-1} \mathcal{D}_t^\top \Omega_t (\mathbf{Y} - \mathbf{X}\theta),$$

where $\Omega_t = \mathrm{diag}\left\{K_h(T_1 - t), ..., K_h(T_n - t)\right\}$ and

$$\mathcal{D}_t = \begin{pmatrix} Z_1^\top & \frac{T_1 - t}{h} Z_1^\top \\ \vdots & \vdots \\ Z_n^\top & \frac{T_n - t}{h} Z_n^\top \end{pmatrix}_{n \times 2r}.$$

Substituting $\{\hat{\alpha}_1(t), ..., \hat{\alpha}_r(t)\}^\top$ in model (2), we obtain

$$\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{X} - \hat{\mathbf{X}})\theta + \varepsilon, \tag{3}$$

where $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$, and

$$\mathbf{S} = \begin{pmatrix} (Z_1^\top\ 0)(\mathcal{D}_{t_1}^\top \Omega_{t_1} \mathcal{D}_{t_1})^{-1} \mathcal{D}_{t_1}^\top \Omega_{t_1} \\ \vdots \\ (Z_n^\top\ 0)(\mathcal{D}_{t_n}^\top \Omega_{t_n} \mathcal{D}_{t_n})^{-1} \mathcal{D}_{t_n}^\top \Omega_{t_n} \end{pmatrix}_{n \times n}.$$

Denote $\widetilde{\mathbf{Y}} = (I_n - \mathbf{S})\mathbf{Y}$ and $\widetilde{\mathbf{X}} = (I_n - \mathbf{S})\mathbf{X}$, where I_n is an $n \times n$ identity matrix. Then model (3) reduces to $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\theta + \varepsilon$, a standard linear regression model for which the ordinary least squares method can be used to estimate θ.

Now, when X_i's are measured with errors, You and Chen (2006) suggested the modified profile least squares estimator

$$\hat{\theta} = \left(\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}} - n\Sigma_u\right)^{-1} \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{Y}}, \tag{4}$$

which is the solution to θ that minimizes

$$\left\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}\theta\right\|^2 - n\theta^\top \Sigma_u \theta,$$

where $\widetilde{\mathbf{W}} = (I_n - \mathbf{S})\mathbf{W}$.

3 Estimation and inference based on model averaging

Unlike You and Chen (2006), who concentrated on the estimation of the coefficients in the linear component of the model based on a single candidate model, we consider estimation based on model averaging. We follow the local mis-specification framework suggested by Hjort and Claeskens (2003) by setting the true value of θ to $\theta_{\text{true}} = (\beta^\top, \gamma_{\text{true}}^\top)^\top = (\beta^\top, \delta^\top/\sqrt{n})^\top$, where the parameter vector $\delta = (\delta_1, ..., \delta_q)^\top$ represents the degree of a model's departure from the narrow model in which $\theta{=}\theta_0{=}(\beta^\top, 0^\top)^\top$. Local parameterization was first introduced by Le Cam (1960), and has been a useful tool for asymptotic analysis.

The results in this section depend on the the following technical conditions, which are also used in Fan and Huang (2005) and You and Chen (2006).

(C1) The random variable T has bounded support $\mathbf{\Omega}$, and its density f is Lipschitz continuous and bounded away from 0 on its support.

(C2) For each $T \in \mathbf{\Omega}$, the $r \times r$ matrix $\mathbf{E}(ZZ^\top|T)$ is non-singular, and $\mathbf{E}(ZZ^\top|T)$, $\mathbf{E}(XX^\top|T)$ and $\mathbf{E}(ZX^\top|T)$ are all Lipschitz continuous.

(C3) There exists some $t > 2$ s.t. $\mathbf{E}\|X\|^{2t} < \infty$, $\mathbf{E}\|Z\|^{2t} < \infty$, $\mathbf{E}\|U\|^{2t} < \infty$ and $\mathbf{E}\|\varepsilon\|^{2t} < \infty$, and $\rho < 2 - t^{-1}$ s.t. $nh^{2\rho-1} \to \infty$.

(C4) $\alpha_j(T), j = 1, ..., r$, is twice continuously differentiable in $T \in \mathbf{\Omega}$.

(C5) $K(\cdot)$ is a symmetric density with compact support.

(C6) The conditions $nh^8 \to 0$ and $nh^2/\{\log(n)\}^2 \to \infty$ are satisfied for the bandwidth h.

Conditions (C1), (C2) and (C4) are related to the degrees of smoothness of the models. Condition (C5) is for the estimator of the unknown function vector to have a closed form expression. Condition (C3) places restrictions on the moments of covariates and bandwidth to guarantee uniform consistency of the kernel estimators. This condition is generally satisfied in practice. For instance, consider $t = 3$, $\rho = 4/3$ and $h = O(n^{-1/5})$. Since $nh^{2\rho-1} = O(n^{2/3}) \to \infty$, all the requirements of condition (C3) are fulfilled. The same bandwidth also satisfies condition (C6) that guarantees the optimal convergence rate of the estimator of the linear component of the model.

3.1 Estimation of coefficients under the full and partially restricted models

When $n \to \infty$, by Lemma A.3 of You and Chen (2006), we have $n^{-1}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}} \xrightarrow{p} \Sigma_u + B$, where $B = \mathbf{E}(X_1X_1^\top) - \mathbf{E}\left[\mathbf{E}(X_1Z_1^\top|T_1)\{\mathbf{E}(Z_1Z_1^\top|T_1)\}^{-1}\mathbf{E}(Z_1X_1^\top|T_1)\right]$, and $\xrightarrow{p}$ denotes convergence in probability. Accordingly, a consistent estimator of B is

$$\hat{B}_n = \frac{1}{n}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}} - \Sigma_u.$$

Partition B, Σ_u, $\widetilde{\mathbf{W}}$, $\widetilde{\mathbf{U}} = (I_n - \mathbf{S})\mathbf{U}$ and $\widetilde{\mathbf{X}}$ conformably with the dimensions of β and γ as $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$, $\Sigma_u = \begin{pmatrix} \Sigma_{u11} & \Sigma_{u12} \\ \Sigma_{u21} & \Sigma_{u22} \end{pmatrix}$, $\widetilde{\mathbf{W}} = \left(\widetilde{\mathbf{W}}_1 | \widetilde{\mathbf{W}}_2\right)$, $\widetilde{\mathbf{U}} = \left(\widetilde{\mathbf{U}}_1 | \widetilde{\mathbf{U}}_2\right)$ and $\widetilde{\mathbf{X}} = \left(\widetilde{\mathbf{X}}_1 | \widetilde{\mathbf{X}}_2\right)$ respectively. Then we can write

$$\begin{pmatrix} \hat{\beta}_{\text{full}} \\ \hat{\gamma}_{\text{full}} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11} & \widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u12} \\ \widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u21} & \widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u22} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{Y}} \\ \widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{Y}} \end{pmatrix}.$$

Direct calculations lead to

$$\hat{\beta}_{\text{full}} = \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11}\right)^{-1} \left\{\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{Y}} - (\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u12})\hat{\gamma}_{\text{full}}\right\} \tag{5}$$

and

$$\hat{\gamma}_{\text{full}} = A_n^{-1} \left\{\widetilde{\mathbf{W}}_2^\top - \left(\widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u21}\right) \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11}\right)^{-1} \widetilde{\mathbf{W}}_1^\top\right\} \widetilde{\mathbf{Y}}, \tag{6}$$

where

$$A_n = \widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u22} - \left(\widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u21}\right) \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11}\right)^{-1} \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u12}\right).$$

Altogether there are 2^q partially restricted models, one for each subset S of $\{1, ..., q\}$; that is, while the partially restricted model includes every element of β, it contains only certain elements of γ_{true}. The full model corresponds to $S = \{1, ..., q\}$, while the narrow model corresponds to $S = \phi$. Denote the coefficients of the partially restricted model in S by β_{S} and γ_{S}. We then have $\beta_{\text{S}} = \beta$ and $\gamma_{\text{S}} = \Pi_{\text{S}}^\top \gamma_{\text{true}}$, where $\Pi_{\text{S}}^\top$ is an $|S| \times q$ selection matrix with the element matching γ_{S} in any given row taking on the value of unity, and zero otherwise, and $|S|$ is the number of components of γ_{true} in the partially restricted model. Similarly, we let $\mathbf{X}_{\text{S}}$, $\widetilde{\mathbf{W}}_{\text{S}}$ and Σ_{us} denote matrices in the partially restricted model S with definitions analogous to the corresponding matrices in the full model. Further, partition these matrices conformably with β_{S} and γ_{S}, and obtain $\mathbf{X}_{\text{S}} = (\mathbf{X}_1 | \mathbf{X}_{2s}) = (\mathbf{X}_1 | \mathbf{X}_2 \Pi_{\text{S}})$, $\widetilde{\mathbf{W}}_{\text{S}} = \left(\widetilde{\mathbf{W}}_1 | \widetilde{\mathbf{W}}_2 \Pi_{\text{S}}\right)$, and $\Sigma_{us} = \begin{pmatrix} \Sigma_{u11} & \Sigma_{u12}\Pi_{\text{S}} \\ \Pi_{\text{S}}^\top \Sigma_{u21} & \Pi_{\text{S}}^\top \Sigma_{u22} \Pi_{\text{S}} \end{pmatrix}$. These manipulations enable the derivation of the following regression coefficient estimators of the partially restricted model in S:

$$\hat{\beta}_{\text{S}} = \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11}\right)^{-1} \left\{\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{Y}} - \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u12}\right) \Pi_{\text{S}} \hat{\gamma}_{\text{S}}\right\} \tag{7}$$

and

$$\hat{\gamma}_{\text{S}} = \left(\Pi_{\text{S}}^\top A_n \Pi_{\text{S}}\right)^{-1} \Pi_{\text{S}}^\top \left\{\widetilde{\mathbf{W}}_2^\top - \left(\widetilde{\mathbf{W}}_2^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u21}\right) \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11}\right)^{-1} \widetilde{\mathbf{W}}_1^\top\right\} \widetilde{\mathbf{Y}}. \tag{8}$$

From equations (5) - (8), we obtain the following set of equations characterizing the relationship between estimators under the full and partially restricted models:

$$\begin{pmatrix} \hat{\beta}_{\text{S}} \\ \hat{\gamma}_{\text{S}} \end{pmatrix} = \begin{pmatrix} I_p & C_{ns} \\ 0_{|S| \times p} & (\Pi_{\text{S}}^\top A_n \Pi_{\text{S}})^{-1} \Pi_{\text{S}}^\top A_n \end{pmatrix} \begin{pmatrix} \hat{\beta}_{\text{full}} \\ \hat{\gamma}_{\text{full}} \end{pmatrix} \equiv G_{ns} \begin{pmatrix} \hat{\beta}_{\text{full}} \\ \hat{\gamma}_{\text{full}} \end{pmatrix}, \tag{9}$$

where $C_{ns} = \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_1 - n\Sigma_{u11}\right)^{-1} \left(\widetilde{\mathbf{W}}_1^\top \widetilde{\mathbf{W}}_2 - n\Sigma_{u12}\right) \left(I_q - A_n^{-1/2} H_{ns} A_n^{1/2}\right)$, and $H_{ns} = A_n^{1/2} \Pi_{\mathrm{S}} \left(\Pi_{\mathrm{S}}^\top A_n \Pi_{\mathrm{S}}\right)^{-1} \Pi_{\mathrm{S}}^\top A_n^{1/2}$.

The following lemma illustrates the asymptotic properties of estimators under the full and restricted models.

Lemma 1. *If conditions (C1)-(C6) hold, and U_i, ε_i and (X_i, Z_i, T_i) are mutually independent, then we have the following convergence result when $n \to \infty$:*

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_{\mathrm{S}} - \beta \\ \hat{\gamma}_{\mathrm{S}}\end{pmatrix} \xrightarrow{d} N\left\{\begin{pmatrix} C_{\mathrm{S}}\delta \\ (\Pi_{\mathrm{S}}^\top A \Pi_{\mathrm{S}})^{-1}\Pi_{\mathrm{S}}^\top A\delta\end{pmatrix},\ G_{\mathrm{S}} P G_{\mathrm{S}}^\top\right\};$$

in particular,

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_{\mathrm{full}} - \beta \\ \hat{\gamma}_{\mathrm{full}}\end{pmatrix} \equiv \begin{pmatrix} M_n \\ \hat{\delta}\end{pmatrix} \xrightarrow{d} \begin{pmatrix} M \\ D\end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ \delta\end{pmatrix},\ P\right\},$$

where $\xrightarrow{d}$ denotes convergence in distribution, $A = B_{22} - B_{21}B_{11}^{-1}B_{12}$, C_{S} and G_{S} are limits of A_n C_{ns} and G_{ns} respectively, $P = B^{-1}FB^{-1}$,

$$F = \mathbf{E}\left(\left[W_i - \mathbf{E}\left(X_i Z_i^\top | T_i\right)\left\{\mathbf{E}\left(Z_i Z_i^\top | T_i\right)\right\}^{-1} Z_i\right]\left(\varepsilon_i - U_i^\top \theta_0\right) + \Sigma_u \theta_0\right)^{\otimes 2},$$

and $R^{\otimes 2} = RR^\top$ for any matrix R.

From the proof of results in the appendix, a consistent estimator of F is

$$\hat{F}_n = \frac{1}{n}\left[\sum_{i=1}^n \left(W_i - \hat{W}_i\right)\left\{\left(Y_i - \hat{Y}_i\right) - \left(W_i - \hat{W}_i\right)^\top \hat{\theta}_{\mathrm{full}}\right\} + \Sigma_u \hat{\theta}_{\mathrm{full}}\right]^{\otimes 2},$$

where $\hat{\theta}_{\mathrm{full}} = (\beta_{\mathrm{full}}^\top, \gamma_{\mathrm{full}}^\top)^\top$. Hence a consistent estimator of the asymptotic variance $G_{\mathrm{S}} P G_{\mathrm{S}}^\top$ is $G_{ns}\hat{P}_n G_{ns}^\top$, where $\hat{P}_n = \hat{B}_n^{-1}\hat{F}_n\hat{B}_n^{-1}$. The bias vector can be estimated by replacing A, C_{S} and δ by A_n, C_{ns} and $\hat{\delta}$ respectively.

3.2 Estimation by model averaging

In this subsection, we consider the estimation of the parameter $\mu_{\mathrm{true}} = \mu(\beta, \gamma_{\mathrm{true}})$ by model averaging. We assume that the parameter of interest μ does not depend on the non-parametric component because the estimator of this component is not $\sqrt{n}$-consistent. Let the estimator based on the partially restricted model in S be $\hat{\mu}_{\mathrm{S}} = \mu(\hat{\beta}_{\mathrm{S}}, \hat{\gamma}_{\mathrm{S}})$. The following theorem can be obtained.

Theorem 1. *Assume that μ is differentiable at $\theta_0 = (\beta^\top, 0^\top)^\top$. If conditions (C1) - (C6) are satisfied, and U_i, ε_i and (X_i, Z_i, T_i) are mutually independent, then we have*

$$\sqrt{n}(\hat{\mu}_{\mathrm{S}} - \mu_{\mathrm{true}}) \xrightarrow{d} \Lambda_{\mathrm{S}} = \mu_\beta^\top\left\{M + B_{11}^{-1}B_{12}(D - \delta)\right\} + \omega^\top\left\{\delta - A^{-1/2}H_{\mathrm{S}}A^{1/2}D\right\}, \qquad (10)$$

where $\omega = B_{21}B_{11}^{-1}\mu_\beta - \mu_\gamma$, $\mu_\beta = \frac{\partial\mu(\beta,0)}{\partial\beta}$, $\mu_\gamma = \frac{\partial\mu(\beta,0)}{\partial\gamma}$, and H_{S} has the same form as H_{ns} except that A_n in H_{ns} is replaced by A in H_{S}.

The asymptotic bias and variance of $\hat{\mu}$ are $\mathbf{E}\Lambda_{\mathrm{S}} = \omega^\top(I_q - A^{-1/2}H_{\mathrm{S}}A^{1/2})\delta$ and $\mathbf{Var}(\Lambda_{\mathrm{S}}) = (\mu_\beta^\top, \mu_\gamma^\top\Pi_{\mathrm{S}})G_{\mathrm{S}}PG_{\mathrm{S}}^\top(\mu_\beta^\top, \mu_\gamma^\top\Pi_{\mathrm{S}})^\top$ respectively. Note that $\mathbf{Var}(\Lambda_{\mathrm{S}})$ can be estimated consistently by using $\hat{\theta}_{\mathrm{full}}$ in μ_β and μ_γ, and replacing G_{S} and P by $\hat{G}_n$ and $\hat{P}_n$ respectively.

With each partially restricted estimator being a submodel estimator, the model average estimator has the form

$$\hat{\mu}_{\mathrm{avg}} = \sum_{\mathrm{S}} c(S|\hat{\delta})\hat{\mu}_{\mathrm{S}}, \tag{11}$$

where $c(S|\hat{\delta})$'s are weight functions that sum to one. Theorem 2 depicts the asymptotic properties of the estimator $\hat{\mu}_{\mathrm{avg}}$.

Theorem 2. *Assume that μ is differentiable at θ_0, and the weight functions $c(S|d)$ are continuous almost everywhere. If conditions (C1) - (C6) hold, and U_i, ε_i and (X_i, Z_i, T_i) are mutually independent, then we have*

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_{\mathrm{avg}} - \mu_{\mathrm{true}}) &\xrightarrow{d} \Lambda = \mu_\beta^\top\left\{M + B_{11}^{-1}B_{12}(D-\delta)\right\} + \omega^\top\left\{\delta - Q(D)D\right\}, \\
\mathbf{E}\Lambda &= \omega^\top\left[\delta - \mathbf{E}\{Q(D)D\}\right], \ \textit{and} \\
\mathbf{Var}(\Lambda) &= \mu_\beta^\top(I, B_{11}^{-1}B_{12})P(I, B_{11}^{-1}B_{12})^\top\mu_\beta + \omega^\top\mathbf{Var}\left\{Q(D)D\right\}\omega \\
&\quad - 2\mu_\beta^\top(I, B_{11}^{-1}B_{12})\mathbf{Cov}\left\{(M^\top, D^\top)^\top, Q(D)D\right\}\omega,
\end{aligned}$$

where $Q(D) = A^{-1/2}\left\{\sum_{\mathrm{S}} c(S|D)H_{\mathrm{S}}\right\}A^{1/2}$. If X were observed without errors, then $M + B_{11}^{-1}B_{12}(D-\delta)$ and D would be independent, and the variance would simplify to

$$\mathbf{Var}(\Lambda) = \mu_\beta^\top B_{11}^{-1}\mu_\beta + \omega^\top\mathbf{Var}\left\{Q(D)D\right\}\omega.$$

This theorem reveals that when there are no measurement errors, the model average estimator under the VCPL model framework has asymptotic mean and variance expressions similar to those of the model average estimators discussed in Hjort and Claeskens (2003, 2006) and Claeskens and Carroll (2007).

3.3 Interval estimation based on model averaging

Note from Theorem 2 that the asymptotic distribution of the model average estimator is non-normal. This concurs with the observation under parametric models in Hjort and Claeskens (2003). Here, we follow Hjort and Claeskens' (2003) approach of constructing confidence interval based on the model average estimator. We demonstrate that the actual coverage probability of the interval converges to the intended level in large samples; as well, we prove that such a confidence interval based on the model average estimator is asymptotically equivalent to that constructed based on the full model estimator that follows an asymptotically normal distribution. The latter result concurs with the findings of Kabaila and Leeb (2006) under a parametric set-up.

Assume that the conditions for Theorem 2 hold. Consider the confidence limits

$$\begin{cases} \mathrm{low}_{\mathrm{avg}} &= \hat{\mu}_{\mathrm{avg}} - \hat{\omega}^\top\left\{\hat{\delta} - Q_n(\hat{\delta})\hat{\delta}\right\}/\sqrt{n} - z\hat{\kappa}/\sqrt{n} \\ \mathrm{up}_{\mathrm{avg}} &= \hat{\mu}_{\mathrm{avg}} - \hat{\omega}^\top\left\{\hat{\delta} - Q_n(\hat{\delta})\hat{\delta}\right\}/\sqrt{n} + z\hat{\kappa}/\sqrt{n}, \end{cases} \tag{12}$$

where z is a standard normal quantile, $\hat{\omega}$ and $\hat{\kappa}$ are consistent estimators of ω and $\kappa = \sqrt{(\mu_\beta^\top, \mu_\gamma^\top) P (\mu_\beta^\top, \mu_\gamma^\top)^\top}$ respectively, and $Q_n(\hat{\delta}) = A_n^{-1/2} \left\{ \sum_{\mathrm{S}} c(S|\hat{\delta}) H_{nS} \right\} A_n^{1/2}$. Then $\Pr\{\mu_{\text{true}} \in (\text{low}_{\text{avg}}, \text{up}_{\text{avg}})\}$=$\Pr\{-z \le T_n \le z\}$ is the probability of the confidence interval containing the true parameter μ_{true}, where

$$T_n = \frac{\sqrt{n}(\hat{\mu}_{\text{avg}} - \mu_{\text{true}}) - \hat{\omega}^\top \left\{ \hat{\delta} - Q_n(\hat{\delta})\hat{\delta} \right\}}{\hat{\kappa}}.$$

As $\sqrt{n}(\hat{\mu}_{\text{avg}} - \mu_{\text{true}})$ is an almost surely continuous function of M_n and $\hat{\delta}$, from the Continuous Mapping Theorem and Slutsky Theorem, we have

$$\left(\sqrt{n}\{\hat{\mu}_{\text{avg}} - \mu_{\text{true}}\}, \hat{\delta}\right) \xrightarrow{d} \left[\Lambda_0 + \omega^\top \{\delta - Q(D)D\}, D\right],$$

where $\Lambda_0 = \mu_\beta^\top \left\{ M + B_{11}^{-1} B_{12}(D - \delta) \right\}$ and $Q(D) = A^{-1/2} \left\{ \sum_{\mathrm{S}} c(S|D) H_{\mathrm{S}} \right\} A^{1/2}$. It follows that

$$T_n \xrightarrow{d} \frac{\Lambda_0 + \omega^\top (\delta - D)}{\kappa} = \frac{\mu_\beta^\top M + \mu_\gamma^\top (D - \delta)}{\kappa}.$$

The limiting variable on the right-hand side of the above equation follows a standard normal distribution. Hence we have $Pr\{-z \le T_n \le z\} \to 2\mathbf{\Phi}(z) - 1$, where $\mathbf{\Phi}$ is the standard normal distribution function.

If we denote $\hat{\mu}_{\text{full}}$ as the estimator of μ under the full model, then by formula (10) in Theorem 1, we obtain the following result:

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{\text{full}} - \mu_{\text{true}}) \xrightarrow{d} & \mu_\beta^\top \left\{ M + B_{11}^{-1} B_{12}(D - \delta) \right\} + \omega^\top (\delta - D) \\ & = \mu_\beta^\top M + \mu_\gamma^\top (D - \delta),\end{aligned}$$

where the limiting variable $\mu_\beta^\top M + \mu_\gamma^\top (D - \delta) \sim N(0, \kappa^2)$. Accordingly, the confidence limits of μ_{true} based on $\hat{\mu}_{\text{full}}$ are

$$\begin{cases} \text{low}_{\text{full}} & = \hat{\mu}_{\text{full}} - z\hat{\kappa}/\sqrt{n} \\ \text{up}_{\text{full}} & = \hat{\mu}_{\text{full}} + z\hat{\kappa}/\sqrt{n}. \end{cases} \tag{13}$$

From the definition of $\hat{\mu}_{\text{avg}}$ and equation (9), and by using the Taylor series expansion, we obtain

$$\begin{aligned}\hat{\mu}_{\text{avg}} &= \mu(\beta, 0) + \sum_{\mathrm{S}} c(S|\hat{\delta}) \begin{pmatrix} \mu_\beta \\ \Pi_{\mathrm{S}}^\top \mu_\gamma \end{pmatrix}^\top \begin{pmatrix} \hat{\beta}_{\mathrm{S}} - \beta \\ \hat{\gamma}_{\mathrm{S}} \end{pmatrix} + o_P(1/\sqrt{n}) \\ &= \mu(\beta, 0) + \mu_\beta^\top (\hat{\beta}_{\text{full}} - \beta) + \sum_{\mathrm{S}} c(S|\hat{\delta}) \left(\mu_\beta^\top C_{ns} \hat{\gamma}_{\text{full}} + \mu_\gamma^\top A_n^{-1/2} H_{ns} A_n^{1/2} \hat{\gamma}_{\text{full}} \right) + o_P(1/\sqrt{n}) \\ &= \mu(\beta, 0) + \mu_\beta^\top (\hat{\beta}_{\text{full}} - \beta) \\ &\quad + \sum_{\mathrm{S}} c(S|\hat{\delta}) \left\{ (\omega^\top + \mu_\gamma^\top)(I_q - A_n^{-1/2} H_{ns} A_n^{1/2}) + \mu_\gamma^\top A_n^{-1/2} H_{ns} A_n^{1/2} \right\} \hat{\delta}/\sqrt{n} + o_P(1/\sqrt{n}) \\ &= \mu(\beta, 0) + \mu_\beta^\top (\hat{\beta}_{\text{full}} - \beta) + \mu_\gamma^\top \hat{\gamma}_{\text{full}} + \omega^\top \{\hat{\delta} - Q_n(\hat{\delta})\hat{\delta}\}/\sqrt{n} + o_P(1/\sqrt{n}).\end{aligned}$$

By the Taylor series expansion,

$$\hat{\mu}_{\text{full}} = \mu(\beta, 0) + \mu_\beta^\top(\hat{\beta}_{\text{full}} - \beta) + \mu_\gamma^\top \hat{\gamma}_{\text{full}} + o_P(1/\sqrt{n}). \tag{14}$$

Therefore,

$$\hat{\mu}_{\text{full}} = \hat{\mu}_{\text{avg}} - \omega^\top\{\hat{\delta} - Q_n(\hat{\delta})\hat{\delta}\}/\sqrt{n} + o_P(1/\sqrt{n}). \tag{15}$$

Comparing equations (12), (13) and (15), we see that $\text{low}_{\text{avg}} = \text{low}_{\text{full}} + o_P(1/\sqrt{n})$ and $\text{up}_{\text{avg}} = \text{up}_{\text{full}} + o_P(1/\sqrt{n})$. Thus, the two confidence intervals, based on the model average estimator and the full model estimator respectively, are asymptotically identical.

More specifically, if μ is a linear combination of β and γ, then the remainder in (14) varnishes. Furthermore, as κ and ω are quantities relevant to the full model only, the estimators $\hat{\kappa}$ and $\hat{\omega}$ are the same for the full model as for the model average. This means if the parameter of interest is a linear combination of regression coefficients, the confidence interval developed based on the model average (i.e., equation (12)) will be exactly identical to that obtained from the full model (i.e., equation (13)). Thus, if the investigator's main concern is interval estimation rather than point estimation, then the confidence interval based on the full model already serves the purpose and model averaging provides no additional useful information. The interval constructed under the full model also has the advantage of being computationally simple.

3.4 Relationship between FMA and model selection estimators

This subsection studies the relationship between the traditional model selection estimators based on information criteria and FMA under the setup of the VCPLE model. Along the lines of Liang and Li (2009), we define the AIC, BIC, and RIC under the VCPLE framework as

$$\begin{aligned}
\text{AIC}_{ns} &= \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}_{\text{S}}\hat{\theta}_{\text{S}}\|^2 - n\hat{\theta}_{\text{S}}^\top \Sigma_{us}\hat{\theta}_{\text{S}} + 2\sigma^2|S|,\\
\text{BIC}_{ns} &= \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}_{\text{S}}\hat{\theta}_{\text{S}}\|^2 - n\hat{\theta}_{\text{S}}^\top \Sigma_{us}\hat{\theta}_{\text{S}} + \sigma^2\log(n)|S|,
\end{aligned}$$

and

$$\text{RIC}_{ns} = \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}_{\text{S}}\hat{\theta}_{\text{S}}\|^2 - n\hat{\theta}_{\text{S}}^\top \Sigma_{us}\hat{\theta}_{\text{S}} + 2\sigma^2\log(|S|)|S|,$$

respectively, where $\hat{\theta}_{\text{S}}$ represents the estimator of regression coefficient in the reduced model. Note that

$$\begin{aligned}
&\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}_{\text{S}}\hat{\theta}_{\text{S}}\|^2 - n\hat{\theta}_{\text{S}}^\top \Sigma_{us}\hat{\theta}_{\text{S}} - \widetilde{\mathbf{Y}}^\top\widetilde{\mathbf{Y}}\\
&= -\hat{\theta}_{\text{S}}^\top(\widetilde{\mathbf{W}}_{\text{S}}^\top\widetilde{\mathbf{W}}_{\text{S}} - n\Sigma_{us})\hat{\theta}_{\text{S}}\\
&= -\begin{pmatrix}\hat{\beta}_{\text{full}}\\ \hat{\gamma}_{\text{full}}\end{pmatrix}^\top G_{ns}^\top \begin{pmatrix} I & 0\\ 0 & \Pi_{\text{S}}^\top\end{pmatrix}\begin{pmatrix}\widetilde{\mathbf{W}}_1^\top\widetilde{\mathbf{W}}_1 - n\Sigma_{u11} & \widetilde{\mathbf{W}}_1^\top\widetilde{\mathbf{W}}_2 - n\Sigma_{u12}\\ \widetilde{\mathbf{W}}_2^\top\widetilde{\mathbf{W}}_1 - n\Sigma_{u21} & \widetilde{\mathbf{W}}_2^\top\widetilde{\mathbf{W}}_2 - n\Sigma_{u22}\end{pmatrix}\begin{pmatrix} I & 0\\ 0 & \Pi_{\text{S}}\end{pmatrix} G_{ns}\begin{pmatrix}\hat{\beta}_{\text{full}}\\ \hat{\gamma}_{\text{full}}\end{pmatrix}\\
&= -\begin{pmatrix}\hat{\beta}_{\text{full}}\\ \hat{\gamma}_{\text{full}}\end{pmatrix}^\top \begin{pmatrix}\widetilde{\mathbf{W}}_1^\top\widetilde{\mathbf{W}}_1 - n\Sigma_{u11} & \widetilde{\mathbf{W}}_1^\top\widetilde{\mathbf{W}}_2 - n\Sigma_{u12}\\ \widetilde{\mathbf{W}}_2^\top\widetilde{\mathbf{W}}_1 - n\Sigma_{u21} & L\end{pmatrix}\begin{pmatrix}\hat{\beta}_{\text{full}}\\ \hat{\gamma}_{\text{full}}\end{pmatrix},
\end{aligned}$$

where $L = -A_n + A_n\Pi_{\rm S}(\Pi_{\rm S}^\top A_n\Pi_{\rm S})^{-1}\Pi_{\rm S}^\top A_n + \widetilde{\mathbf{W}}_2^\top\widetilde{\mathbf{W}}_2 - n\Sigma_{u22}$. Note that in the last equality, only $A_n\Pi_{\rm S}(\Pi_{\rm S}^\top A_n\Pi_{\rm S})^{-1}\Pi_{\rm S}^\top A_n$ depends on S. In addition, $A_n/n \xrightarrow{p} A$ as $n\to\infty$. Hence, we have

$$\begin{aligned}
&\left\{\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}_{\rm S}\hat\theta_{\rm S}\|^2 - n\hat\theta_{\rm S}^\top\Sigma_{us}\hat\theta_{\rm S}\right\} - \left\{\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{W}}_\phi\hat\theta_\phi\|^2 - n\hat\theta_\phi^\top\Sigma_{u\phi}\hat\theta_\phi\right\}\\
&= -\begin{pmatrix}\hat\beta_{\rm full}\\ \hat\gamma_{\rm full}\end{pmatrix}^\top \begin{pmatrix}0 & 0\\ 0 & A_n\Pi_{\rm S}(\Pi_{\rm S}^\top A_n\Pi_{\rm S})^{-1}\Pi_{\rm S}^\top A_n\end{pmatrix}\begin{pmatrix}\hat\beta_{\rm full}\\ \hat\gamma_{\rm full}\end{pmatrix}\\
&= -\hat\gamma^\top A_n\Pi_{\rm S}(\Pi_{\rm S}^\top A_n\Pi_{\rm S})^{-1}\Pi_{\rm S}^\top A_n\hat\gamma\\
&= -\hat\delta^\top A\Pi_{\rm S}(\Pi_{\rm S}^\top A\Pi_{\rm S})^{-1}\Pi_{\rm S}^\top A\hat\delta + o_P(1).
\end{aligned}$$

This implies that given the estimator $\hat\gamma$ under the full model, the relative magnitudes of an information criterion (say, the AIC) across different submodels are determined by the selection matrix which is a function of the set S. Therefore, asymptotically, the AIC model selection estimator can be viewed as a model average estimator in the form of equation (11) with indicator functions as its weights. For example, assuming that there are no ties among the AIC values, the AIC model selection estimator can be written as

$$\begin{aligned}
\hat\mu_{\rm AIC} &= \sum_{\rm S} I_{\{{\rm AIC_S\ is\ the\ smallest}\}}\hat\mu_{\rm S}\\
&= \sum_{\rm S} I_{\{-\hat\delta^\top A\Pi_{\rm S}(\Pi_{\rm S}^\top A\Pi_{\rm S})^{-1}\Pi_{\rm S}^\top A\hat\delta + 2\sigma^2|S|\ {\rm is\ the\ smallest}\}}\hat\mu_{\rm S}, \text{ for a large } n\\
&\equiv \sum_{\rm S} c(S|\hat\delta)\hat\mu_{\rm S}.
\end{aligned}$$

The same result holds for the BIC and RIC model selection criteria. Evidently, the variance of $\hat\mu_{\rm AIC}$ differs from the variance of $\hat\mu_S$ for each set S because the indicator function is also random. However, typically, the investigator uses the variance of the estimator from the chosen model (i.e., $\mathbf{Var}(\hat\mu_S) = \mathbf{Var}(\hat\mu_{\rm AIC}|{\rm AIC}_S$ is the smallest)). We call this approach the naive approach to distinguish it from the post selection approach of Hjort and Claeskens (2003), by which the variation of the indicator function is also taken into account[1].

4 Finite sample analysis by simulations

In this section, we evaluate the finite sample performance of the FMA estimator through simulations. The implementation of our method requires the selection of bandwidth for the non-parametric component of the model. This is an important yet unsolved problem for semi-parametric modeling (Fan and Huang, 2005). We will not elaborate upon this problem here, because we focus primarily on the estimation of parameters in the linear component of the model, which is insensitive to the choice of the bandwidth. In our simulations we use a cross-validation method to choose the bandwidth parameter.

Our simulation study is based on the model

$$\begin{aligned}
Y &= X^\top\theta + Z_1\sin(2\pi T) + Z_2\sin(6\pi T) + \varepsilon,\\
W &= X + U
\end{aligned} \tag{16}$$

[1]The point estimates are the same under both approaches.

where $X = (X_1, X_2, X_3, X_4, X_5)^\top$, $\theta = \{\beta^\top, \gamma^\top\}^\top = \{(1.5, 2), \delta^\top/\sqrt{n}\}^\top$, $X_1, ..., X_5$, Z_1, and Z_2 are covariates, each having a standard normal distribution with ρ being the correlation coefficient of each pair of covariates, $T \sim \text{Uniform}(0, 1)$, $U \sim N(0,\ \sigma_u^2 I)$, $\varepsilon \sim N(0,\ 1)$, $\sigma_u = 0.1$, 0.5, and we let δ be $\delta^{(1)} = (0, 0, 0)^\top$, $\delta^{(2)} = (1, 0, 1)^\top$ and $\delta^{(3)} = (1, 1, 1)^\top$. We focus our interest on the following three estimands: $\mu_1 = \beta_1 - \beta_2 + \gamma_3$, $\mu_2 = \beta_1 + \beta_2 + \gamma_1 + \gamma_2 + \gamma_3$, and $\mu_3 = 5\beta_2/(\beta_1 + \beta_2 + \gamma_1 + \gamma_2 + \gamma_3)$. These estimands are reflective of common situations in practice where interest often centers on linear and non-linear combinations of parameters. For example, in economics, the estimate of returns to scale in a Cobb-Douglas production function is provided by the sum of the coefficient estimates; also, in demand analysis, long-run elasticities of dynamic models can be defined as a non-linear function of the estimated parameters (Hirschberg *et al.*, 2008).

In constructing the model average, we follow Buckland et al.'s (1997) suggestion of assigning the weights based on smoothed AIC (S-AIC) and smoothed BIC (S-BIC) values, represented by $\frac{\exp(-\frac{1}{2}\text{AIC}_{ns})}{\sum_{\text{S}} \exp(-\frac{1}{2}\text{AIC}_{ns})}$ and $\frac{\exp(-\frac{1}{2}\text{BIC}_{ns})}{\sum_{\text{S}} \exp(-\frac{1}{2}\text{BIC}_{ns})}$ respectively. We compare the FMA estimator with estimators from the full model along with AIC- and BIC-based model selection, and the performance of estimators is evaluated in terms of MSE. In each case we draw $R = 1000$ independent samples of size n, and the MSE of the estimator of μ_i is calculated based on the formula $\text{MSE} = R^{-1} \sum_{r=1}^{R} (\hat{\mu}_i^{(r)} - \mu_i)^2, i = 1, 2, 3$, where $\hat{\mu}_i^{(r)}$ is the estimate of μ_i obtained from the r-th run.

We refer to the ratio of the MSE for a given method to the MSE of the full model estimator as relative MSE (RMSE). Thus, a RMSE smaller than unity indicates that the given method is superior to the full model estimator, and vice versa. Tables 1 and 2 report the results for $\rho = 0.5$ and $\rho = 0$ respectively[2]. To facilitate readability, the smallest RMSE in each panel is flagged by a "†". Note that the full model estimator is always unbiased even if over-fits the true model, but its variance can be larger than those produced by estimators that are biased.

The following observations may be noted from the results. First, of all cases considered, model averaging invariably delivers superior RMSE than its model selection counterpart. Although there are exceptions, this superiority is generally more marked when $\sigma_u = 0.1$ than when $\sigma_u = 0.5$, and when $\rho = 0$ than when $\rho = 0.5$, *ceteris paribus*. Second, in the case of $\delta = \delta^{(1)}$, no matter the values of ρ, σ_u and n, the full model estimator always yields the worst estimates; this is hardly surprising as the full model is grossly over-fitted when $\delta = \delta^{(1)}$. For the other two choices of δ, full model estimation, with few exceptions, remains inferior to model averaging, but it can be a better strategy than model selection in a good number of cases. The improved performance of the full model estimator for these choices of δ is of no surprise - when $\delta = \delta^{(2)}$ or $\delta = \delta^{(3)}$, the full model is either only mildly over-fitted or correctly specified. However, although the full model estimator is always asymptotically unbiased, in most cases the variance produced by the full model estimator remains larger than those produced by other strategies. Thus, the full model estimator frequently remains worse than the other estimators even when the full model is the true model or close to being the true model. That being said, the AIC and BIC model selection estimators both perform poorly when $\rho = 0$, $\delta = \delta^{(3)}$, and $\sigma_u = 0.1$, having MSEs that are larger than that of the full

[2]We have also considered covariance structures of explanatory variables other than those described here but the results are not reported due to their similarity to those presented here.

model estimator for all three estimands and both values of n. Interestingly, it is also under these choices of ρ, δ and σ_u that model averaging is sometimes found to perform worse than full model estimation. This can be partially explained by noting that when $\delta = \delta^{(3)}$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$, and with the majority of submodels in the model average having at least one $\gamma_j = 0$, the model average estimator will likely be substantially biased. When $\sigma_u = 0.5$, model selection typically has an edge over full model estimation irrespective of δ. Third, in the large majority of cases considered, the S-BIC model average estimator yields the smallest MSE; in the remaining cases where S-BIC model averaging is not the best strategy, the most accurate estimates are invariably produced by S-AIC averaging. In other words, for all of the cases considered, the dominating estimator is always either the S-AIC or the S-BIC model average estimator. Fourth, of the two model selection estimators, the BIC estimator is generally preferred to the AIC estimator, and there are a good number of instances where the BIC model selection estimator has smaller MSE than the S-AIC model average estimator. Commonly, the RMSE comparisons of estimators for $n = 100$ and 200 are reasonably similar.

5 Analysis of real data

Here, we apply our method to a subset of data obtained from the *Continuing Survey of Food Intakes by Individuals* (CSFII) conducted by the U.S. Department of Agriculture in 1985 and 1986. [3]

This data set contains dietary intake and related information of $n = 1827$ individuals between the age of 25 and 50. Using the available data, we specify the following model for calorie intake, denoted by y:

$$y = \sum_{i=1}^{7} \beta_i x_i + f_0(t) + z f_1(t) + \varepsilon,$$

where x_1, x_2, x_3, x_4 and x_5 represent intake levels of fat, protein, carbohydrates, Vitamin A and Vitamin C respectively, x_6 is an indicator variable for alcohol consumption, x_7 is body mass index, z is income and t is age. As we think that fat, protein and carbohydrates are the key determinants of calories, and we are primarily interested in the effects that these variables have on calorie intake, we treat x_1, x_2 and x_3 as mandatory in the parametric component of the model. Indeed, statistical results based on the full model reveal that only x_1, x_2, x_3 and x_6 are significant, and the coefficient estimates (in absolute values) of x_1, x_2 and x_3 are at least three-fold those of the other variables. As we are less interested in the effects of x_4, x_5, x_6, and x_7 on y, we treat this second group of variables as optional. One key role of the optional variables is to improve the estimation of the coefficients of the mandatory variables. This approach for distinguishing between mandatory and optional explanatory variables is adopted from Magnus and Durbin (1999) and Danilov and Magnus (2004).

We are interested in the estimation of the following four estimands: $\mu_1 = \beta_1$, $\mu_2 = \beta_2$, $\mu_3 = \beta_3$ and $\mu_4 = \beta_1/\beta_2$, based on five alternative estimation methods: FMA by S-AIC and

[3] This is part of the Nationwide Food Consumption Survey, published by the U.S. Department of Agriculture Human Nutrition Information Service, Hyattsville, Maryland, CSFII Reports No. 85-4 and No. 86-3.

S-BIC, model selection by AIC and BIC, and full model estimation. The estimands μ_1, μ_2 and μ_3 are of obvious interest because they represent the marginal effects that each of the mandatory explanatory variables have on calorie intake. The estimand μ_4 is also of interest as it measures the effect of fat relative to that of protein. Tables 3 and 4 present the point and interval estimation results. We observe from the tables that the two model selection methods produce identical results - for this data set, both the AIC and BIC select the model that contains x_1, x_2, x_3, x_4 and x_6. Results produced by the two FMA estimators are also quite similar; both S-AIC and S-BIC model averaging yield estimates of μ_1 and μ_4 that are larger and estimates of μ_2 and μ_3 that are smaller than the corresponding estimates obtained from model selection and full model estimation. The relatively large μ_4 estimates produced by the two model average estimators indicate that among the estimation approaches considered, model averaging most accentuates the common belief that calorie intake is associated with fat consumption more than with protein consumption. As for interval estimation comparisons, note that for μ_1, μ_2 and μ_3, model averaging and the full model estimation produce the same interval estimates as these estimands are all linear in parameters. Table 4 shows that model selection generally results in wider confidence intervals than do full model estimation or model averaging.

6 Concluding remarks

In this section, we summarize our main findings and point to some directions for future research.

- In the context of the VCPLE model, we have considered frequentist model averaging in the manner of Hjort and Claeskens (2003). We have derived the asymptotic distribution of the FMA estimator of the unknown parameters of interest, and developed a confidence interval procedure based on the FMA estimator. Asymptotically, the resultant interval achieves the target nominal coverage probability, and is identical to the confidence interval obtained from the estimation of the full model. More remarkably, if the parameter of interest is a linear combination of regression coefficients, then the equivalence between the FMA and full model based confidence intervals also holds in finite samples. In view of the simulation findings suggesting that FMA generally has an advantage over full model estimation in point estimation, alternative methods of interval estimation based on the FMA approach resulting in more efficient estimates likely exist, and this is an area that undoubtedly deserves more study.

- Throughout this paper, we assume that Σ_u is known. To estimate Σ_u when it is unknown, it is usually assumed that replicated observations of X_i are available such that $W_{ij} = X_i + U_{ij}$, $j = 1, ..., J_i$, $i = 1, ..., n$ are observed (Carroll *et al.*, 2006; Liang and Li, 2009). Then Σ_u can be consistently and unbiasedly estimated by

$$\hat{\Sigma}_u = \frac{\sum_{i=1}^{n} \sum_{j=1}^{J_i} (W_{ij} - \bar{W}_i)(W_{ij} - \bar{W}_i)^\top}{\sum_{i=1}^{n} (J_i - 1)},$$

  where $\bar{W}_i = \sum_{j=1}^{J_i} W_{ij}/J_i$. The substitution of Σ_u by $\hat{\Sigma}_u$ does not complicate the theoretical analysis in any substantial way; all asymptotic results continue to hold when

Σ_u is replaced by a consistent estimator. Having said that, since $\bar{U}_i = \sum_{j=1}^{J_i} U_{ij}/J_i$ has smaller variance than U_{ij}, an arguably better way to proceed would be to modify model (1) as

$$\begin{cases} Y_i &= X_i^\top \theta + Z_i^\top \alpha(T_i) + \varepsilon_i \\ \bar{W}_i &= X_i + \bar{U}_i. \end{cases}$$

In this case, the distributions of $\bar{U}_i$'s are different if J_i's are not all identical. Then the expression of F in Lemma 1 should be modified to the following, assuming that the limit exists:

$$F = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left(\left[\bar{W}_i - \mathbf{E}(X_i Z_i^\top | T_i) \left\{ \mathbf{E}(Z_i Z_i^\top | T_i) \right\}^{-1} Z_i \right] \left(\varepsilon_i - \bar{U}_i^\top \theta_0 \right) + \Sigma_u \theta_0 \right)^{\otimes 2}$$

- The only type of measurement errors we have considered is one where the errors are present in the linear part of the model. Cases where covariates in the non-parametric part are measured with errors, or measurement errors arise in a more general framework, such as the generalized varying-coefficient partially linear model $g(\mathbf{E}Y_i) = X_i^\top \theta + Z_i^\top \alpha(T_i)$, are definitely worthy of study.

- While we considered only model averaging based on weights constructed from values of AIC and BIC, other weight choice techniques exist (Hansen, 2007, 2008, 2010; Wan *et al.*, 2010; Hansen and Racine, 2012). The consideration of these alternative weight choice mechanisms in the context of the VCPLE model also warrants future studies.

- While we assumed i.i.d. observations, the extension to the non-i.i.d. situation will be a fruitful avenue for future research. Wang and Zou (2012) recently considered model averaging with non-i.i.d. observations in a linear measurement error model, which is a special case of the more general VCPLE model framework examined here.

- It should be mentioned that although the FMA strategy being studied produces a $\sqrt{n}$-consistent estimator of the parametric component of the model, this strategy when applied to the non-parametric component of the model does not yield an estimator that converges to the unknown function at the rate of $1/\sqrt{n}$. It is for this reason that throughout the paper we focused only on the estimators in the parametric component. It remains for future research to develop an FMA strategy for the non-parametric component that possesses optimal properties.

Acknowledgements We thank the editor, associate editor, referees and Professor Ejaz Ahmed for their helpful comments and suggestions. Zou's work was supported by the National Natural Science Foundation of China (Grant nos. 11021161, 70933003 and 70625004) and the Hundred Talents Program of the Chinese Academy of Sciences. Wan's work was supported by a General Research Fund from the Hong Kong Research Grants Council (Grant no. CityU-102709). An earlier version of this paper was presented at the 3rd International Conference of the ERCIM Working Group on Computing and Statistics, London, December 2010.

7 Appendix

Proof of Lemma 1. Write

$$\hat{W}_i = [(Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{W}]^\top, \qquad \hat{U}_i = [(Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{U}]^\top,$$
$$\hat{Y}_i = (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{Y}, \qquad \text{and } \hat{\varepsilon}_i = (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \boldsymbol{\varepsilon},$$

and let $\bigtriangledown = \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}} - n\Sigma_u$. Then from equation (4), we have

$$\begin{aligned}
\hat{\theta} - \theta_{\text{true}} &= \bigtriangledown^{-1} \sum_{i=1}^n (W_i - \hat{W}_i)(Y_i - \hat{Y}_i) - \bigtriangledown^{-1} \bigtriangledown \theta_{\text{true}} \\
&= \bigtriangledown^{-1} \Big[\sum_{i=1}^n (W_i - \hat{W}_i)(Y_i - \hat{Y}_i) - \sum_{i=1}^n (W_i - \hat{W}_i)(W_i - \hat{W}_i)^\top \theta_{\text{true}} + n\Sigma_u \theta_{\text{true}} \Big] \\
&= \bigtriangledown^{-1} n\Sigma_u \theta_{\text{true}} + \bigtriangledown^{-1} \sum_{i=1}^n (W_i - \hat{W}_i) \Big[Y_i - \hat{Y}_i - (W_i - \hat{W}_i)^\top \theta_{\text{true}} \Big].
\end{aligned}$$

From the expressions of $\hat{Y}_i$, $\hat{W}_i$, $\hat{U}_i$ and $\hat{\varepsilon}_i$, we obtain

$$\begin{aligned}
&Y_i - \hat{Y}_i - (W_i - \hat{W}_i)^\top \theta_{\text{true}} \\
=&X_i^\top \theta_{\text{true}} + Z_i^\top \alpha(T_i) + \varepsilon_i - (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{Y} - X_i^\top \theta_{\text{true}} - U_i^\top \theta_{\text{true}} \\
&+ (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{X} \theta_{\text{true}} + (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{U} \theta_{\text{true}} \\
=&Z_i^\top \alpha(T_i) + \varepsilon_i - U_i^\top \theta_{\text{true}} - \hat{\varepsilon}_i + \hat{U}_i^\top \theta_{\text{true}} - (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{M}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^n (W_i - \hat{W}_i) \Big(Y_i - \hat{Y}_i - (W_i - \hat{W}_i) \theta_{\text{true}} \Big) \\
=&\sum_{i=1}^n (W_i - \hat{W}_i)(\varepsilon_i - U_i^\top \theta_{\text{true}}) + \sum_{i=1}^n (W_i - \hat{W}_i)(\hat{U}_i^\top \theta_{\text{true}} - \hat{\varepsilon}_i) \\
&+ \sum_{i=1}^n (W_i - \hat{W}_i) \{ Z_i^\top \alpha(T_i) - (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{M} \} \\
=&\sum_{i=1}^n [W_i - \mathbf{E}(W_i Z_i^\top | T_i) \{ \mathbf{E}(Z_i Z_i^\top | T_i) \}^{-1} Z_i](\varepsilon_i - U_i^\top \theta_{\text{true}}) \\
&+ \sum_{i=1}^n [\mathbf{E}(W_i Z_i^\top | T_i) \{ \mathbf{E}(Z_i Z_i^\top | T_i) \}^{-1} Z_i - \hat{W}_i](\varepsilon_i - U_i^\top \theta_{\text{true}}) \\
&+ \sum_{i=1}^n (W_i - \hat{W}_i)(\hat{U}_i^\top \theta_{\text{true}} - \hat{\varepsilon}_i) \\
&+ \sum_{i=1}^n (W_i - \hat{W}_i) \{ Z_i^\top \alpha(T_i) - (Z_i^\top\ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{M} \} \\
\equiv&J_1 + J_2 + J_3 + J_4.
\end{aligned}$$

Note that J_1 is a sum of the i.i.d. random vectors, and we can easily show that it achieves asymptotic normality by applying the Central Limit Theorem. We now show that each of J_2, J_3, and J_4 is of order $o_P(\sqrt{n})$. Following the method used in Fan and Huang (2005), we can show that the following equation holds uniformly in T:

$$(Z_i^\top \ 0)(\mathcal{D}_{t_i}^\top \Omega_{t_i} \mathcal{D}_{t_i})^{-1} \mathcal{D}_{t_i}^\top \Omega_{t_i} \mathbf{W} = Z_i^\top \{\mathbf{E}(Z_i Z_i^\top | T_i)\}^{-1} \mathbf{E}(Z_i X_i^\top | T_i)\{1 + O_P(c_n)\},$$

where $c_n = \sqrt{\frac{\log(1/h)}{nh}} + h^2$. In addition, since $[\mathbf{E}(W_i Z_i^\top | T_i)]^\top = \mathbf{E}(Z_i W_i^\top | T_i)$ and $\theta_{\text{true}} = \theta_0 + (0^\top, \delta^\top)^\top / \sqrt{n}$, we have

$$J_2 = \sum_{i=1}^n [\mathbf{E}(W_i Z_i^\top | T_i)\{\mathbf{E}(Z_i Z_i^\top | T_i)\}^{-1} Z_i](\varepsilon_i - U_i^\top \theta_0) O_P(c_n).$$

The application of the Central Limit Theorem yields $\sum_{i=1}^n [\mathbf{E}(W_i Z_i^\top | T_i)\{\mathbf{E}(Z_i Z_i^\top | T_i)\}^{-1} Z_i](\varepsilon_i - U_i^\top \theta_0) = O_P(\sqrt{n})$. Therefore, $J_2 = O_P(\sqrt{n} c_n) = o_P(\sqrt{n})$. Similarly, we can show that $J_3 = o_P(\sqrt{n})$, and $J_4 = o_P(\sqrt{n})$. Using the Slutsky Theorem and recognizing that $\bigtriangledown / n = B_n \xrightarrow{p} B$ as $n \to \infty$, we obtain

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_{\text{true}}) =& \left(\frac{\bigtriangledown}{n}\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \Bigg\{ \Big(W_i - E(W_i Z_i^\top | T_i)\{E(Z_i Z_i^\top | T_i)\}^{-1} Z_i \Big) \\
& \times \left[\varepsilon_i - U_i^\top (\beta^\top, \delta^\top / \sqrt{n})^\top\right] + \Sigma_u (\beta^\top, \delta^\top / \sqrt{n})^\top \Bigg\} + o_P(1) \\
=& \left(\frac{\bigtriangledown}{n}\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \Bigg\{ \Big(W_i - E(W_i Z_i^\top | T_i)\{E(Z_i Z_i^\top | T_i)\}^{-1} Z_i \Big) \\
& \times (\varepsilon_i - U_i^\top \theta_0) + \Sigma_u \theta_0 \Bigg\} + o_P(1) \\
\xrightarrow{d} & N(0, \ B^{-1} F B^{-1}).
\end{aligned}$$

By the Continuous Mapping Theorem, $G_{ns} \xrightarrow{p} G_{\text{S}}$. Applying the Continuous Mapping Theorem again and using the above results, we have

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_{\text{S}} - \beta \\ \hat{\gamma}_{\text{S}} \end{pmatrix} \xrightarrow{d} G_{\text{S}} \begin{pmatrix} M \\ D \end{pmatrix} \sim N\left(\begin{pmatrix} C_{\text{S}} \delta \\ (\Pi_{\text{S}}^\top A \Pi_{\text{S}})^{-1} \Pi_{\text{S}}^\top A \delta \end{pmatrix}, \ G_{\text{S}} B^{-1} F B^{-1} G_{\text{S}}^\top \right).$$

□

Proof of Theorem 1. By the Taylor series expansion, we have

$$\mu_{\text{true}} = \mu\left(\beta, \frac{\delta}{\sqrt{n}}\right) = \mu(\beta, 0) + \mu_\gamma^\top \frac{\delta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\hat{\mu}_{\text{S}} = \mu(\hat{\beta}_{\text{S}}, \hat{\gamma}_{\text{S}}) = \mu(\beta, 0) + \begin{pmatrix} \mu_\beta \\ \Pi_{\text{S}}^\top \mu_\gamma \end{pmatrix}^\top \begin{pmatrix} \hat{\beta}_{\text{S}} - \beta \\ \hat{\gamma}_{\text{S}} \end{pmatrix} + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{17}$$

Hence,

$$\begin{aligned}\hat{\mu}_{\mathrm{S}} - \mu_{\mathrm{true}} &= \begin{pmatrix} \mu_\beta \\ \Pi_{\mathrm{S}}^\top \mu_\gamma \end{pmatrix}^\top \begin{pmatrix} \hat{\beta}_{\mathrm{S}} - \beta \\ \hat{\gamma}_{\mathrm{S}} \end{pmatrix} - \mu_\gamma^\top \frac{\delta}{\sqrt{n}} + o_P\left(\frac{1}{\sqrt{n}}\right) \\ &= \begin{pmatrix} \mu_\beta \\ \Pi_{\mathrm{S}}^\top \mu_\gamma \end{pmatrix}^\top G_{ns} \begin{pmatrix} \hat{\beta}_{\mathrm{full}} - \beta \\ \hat{\gamma}_{\mathrm{full}} \end{pmatrix} - \mu_\gamma^\top \frac{\delta}{\sqrt{n}} + o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Note that $M_n = \sqrt{n}(\hat{\beta}_{\mathrm{full}} - \beta)$ and $\hat{\delta} = \sqrt{n}\hat{\gamma}_{\mathrm{full}}$. As

$$G_{ns} = \begin{pmatrix} I_p & C_{ns} \\ 0_{|S|\times p} & (\Pi_{\mathrm{S}}^\top A_n \Pi_{\mathrm{S}})^{-1}\Pi_{\mathrm{S}}^\top A_n \end{pmatrix},$$

it can be shown that

$$\sqrt{n}(\hat{\mu}_{\mathrm{S}} - \mu_{\mathrm{true}}) = \mu_\beta^\top [M_n + B_{11}^{-1}B_{12}(\hat{\delta} - \delta)] + \omega^\top \Big(\delta - A_n^{-1/2} H_{\mathrm{S}} A_n^{1/2}\hat{\delta}\Big) + o_P(1).$$

From the Continuous Mapping Theorem and Slutsky Theorem, the above equation converges in distribution to the following variable:

$$\Lambda_{\mathrm{S}} = \mu_\beta^\top [M + B_{11}^{-1}B_{12}(D - \delta)] + \omega^\top \Big(\delta - A^{-1/2} H_{\mathrm{S}} A^{1/2} D\Big).$$

□

Proof of Theorem 2. From the definition of the FMA estimator $\hat{\mu}_{\mathrm{avg}}$ in equation (11), we have

$$\sqrt{n}(\hat{\mu}_{\mathrm{avg}} - \mu_{\mathrm{true}}) = \sum_{\mathrm{S}} c(S|\hat{\delta})\sqrt{n}(\hat{\mu}_{\mathrm{S}} - \mu_{\mathrm{true}}). \tag{18}$$

From the proof of Theorem 1, $\sqrt{n}(\hat{\mu}_{\mathrm{S}} - \mu_{\mathrm{true}})$ on the right-hand side of (18) can be represented by a linear function of M_n and $\hat{\delta}$. As $c(S|d)$ is almost surely continuous, $\sqrt{n}(\hat{\mu} - \mu_{\mathrm{true}})$ is an almost surely continuous function of M_n and $\hat{\delta}$. Thus, applying the Continuous Mapping Theorem, Slutsky Theorem, and Theorem 1, we obtain the required result. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory,* B.N. Petrov and F. Csaki (eds.), 267–281, Akademiai Kaido: Budapest.

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

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, 2nd Edition.* Chapman and Hall: New York.

Claeskens, G. and Carroll, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249–265.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98**, 900–916.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, New York.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 1, 81–94.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics* **122**, 27–46.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society (Series B)* **57**, 45–97.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310–320.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031–1057.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* **22**, 1947–1975.

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

Hansen, B. E. (2008). Least squares forecast averaging. *Journal of Econometrics* **146**, 342–350.

Hansen, B. E. (2010). Averaging estimators for autoregressions with a near unit root. *Journal of Econometrics* **158**, 142–155.

Hansen, B. E. and Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics* **167**, 38–46.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, (Series B)* **55**, 757–796.

Hirschberg, J., Lye, J., and Slottje, D. (2008). Inferential methods for elasticity estimates. *Journal of Econometrics* **147**, 299–315.

Hjort, N. L. and Claeskens, G. (2003). Frequestist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the cox hazard regression model. *Journal of the American Statistical Association* **101**, 1449–1464.

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

Kabaila, P. and Leeb, H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* **101**, 619–629.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics* **3**, 37–98.

Leung, G. and Barron, A. (2006). Infromation theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52**, 3396–3410.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association* **104**, 234–248.

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* **106**, 1053–1066.

Magnus, J. and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* **67**, 639–643.

Mallows, C. (1973). Some comments on cp. *Technometrics* **15**, 661–675.

Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)* **36**, 111–147.

Wan, A. T. K., Zhang, X., and Zou, G. (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics* **156**, 277–283.

Wang, H., Zhang, X., and Zou, G. (2009). Frequentist model averaging estimation: A review. *Journal of Systems Science and Complexity* **22**, 732–748.

Wang, H. and Zou, G. (2012). Frequentist model average estimation for linear errors-in-variables model. *Journal of Systems Science and Mathematical Science* **32**, 1–14.

Yang, Y. (2003). Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* **13**, 783–809.

You, J. and Chen, G. (2006). Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model. *Journal of Multivariate Analysis* **97**, 324–341.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association* **100**, 1202–1214.

Zhang, W., Lee, S. Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis* **82**, 166–188.

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

Zhang, X., Wan, A. T. K., and Zhou, S. (2012). Focused information criteria, model selection and model averaging in a tobit model with a non-zero threshold. *Journal of Business and Economic Statistics* **30**, 132–142.

Table 1: RMSE of estimators when $\rho = 0.5$

		$\delta^{(1)}$		$\delta^{(2)}$		$\delta^{(3)}$	
	$\sigma_u =$	0.1	0.5	0.1	0.5	0.1	0.5
n=100							
μ_1	S-AIC averaging	0.854	0.947	0.912	0.956	0.913	0.956
	AIC selection	0.924	0.986	0.985	0.982	0.984	0.989
	S-BIC averaging	$0.786^{\dagger}$	$0.902^{\dagger}$	$0.872^{\dagger}$	$0.920^{\dagger}$	$0.876^{\dagger}$	$0.921^{\dagger}$
	BIC selection	0.825	0.927	0.955	0.955	0.979	0.951
μ_2	S-AIC averaging	0.854	0.960	0.904	0.959	0.929	0.956
	AIC selection	0.939	0.998	0.978	0.988	1.014	0.987
	S-BIC averaging	$0.761^{\dagger}$	$0.926^{\dagger}$	$0.857^{\dagger}$	$0.928^{\dagger}$	$0.919^{\dagger}$	$0.922^{\dagger}$
	BIC selection	0.809	0.961	0.949	0.956	1.028	0.966
μ_3	S-AIC averaging	0.800	0.934	0.884	0.945	$0.933^{\dagger}$	0.950
	AIC selection	0.900	0.992	0.955	0.984	1.004	0.986
	S-BIC averaging	$0.677^{\dagger}$	$0.873^{\dagger}$	$0.835^{\dagger}$	$0.902^{\dagger}$	0.946	$0.914^{\dagger}$
	BIC selection	0.745	0.901	0.950	0.935	1.100	0.949
n=200							
μ_1	S-AIC averaging	0.856	0.937	0.914	0.949	0.916	0.951
	AIC selection	0.935	0.966	0.995	0.991	0.977	0.998
	S-BIC averaging	$0.778^{\dagger}$	$0.879^{\dagger}$	$0.882^{\dagger}$	$0.906^{\dagger}$	$0.887^{\dagger}$	$0.907^{\dagger}$
	BIC selection	0.795	0.920	0.956	0.945	0.970	0.942
μ_2	S-AIC averaging	0.850	0.954	0.894	0.958	$0.932^{\dagger}$	0.954
	AIC selection	0.921	0.987	0.966	0.983	0.991	0.983
	S-BIC averaging	$0.747^{\dagger}$	$0.914^{\dagger}$	$0.853^{\dagger}$	$0.918^{\dagger}$	0.958	$0.913^{\dagger}$
	BIC selection	0.783	0.951	0.914	0.955	1.044	0.951
μ_3	S-AIC averaging	0.770	0.935	0.868	0.952	$0.928^{\dagger}$	0.953
	AIC selection	0.881	0.978	0.966	1.001	1.011	0.990
	S-BIC averaging	$0.611^{\dagger}$	$0.867^{\dagger}$	$0.815^{\dagger}$	$0.905^{\dagger}$	0.985	$0.920^{\dagger}$
	BIC selection	0.653	0.916	0.920	0.944	1.143	0.969

Table 2: RMSE of estimators when $\rho = 0$

		$\delta^{(1)}$		$\delta^{(2)}$		$\delta^{(3)}$	
	$\sigma_u =$	0.1	0.5	0.1	0.5	0.1	0.5
n=100							
μ_1	S-AIC averaging	0.844	0.937	0.934	0.961	0.931	0.961
	AIC selection	0.920	0.978	1.012	0.998	1.017	0.997
	S-BIC averaging	$0.766^{\dagger}$	$0.889^{\dagger}$	$0.905^{\dagger}$	$0.930^{\dagger}$	$0.900^{\dagger}$	$0.931^{\dagger}$
	BIC selection	0.803	0.939	1.008	0.969	1.013	0.976
μ_2	S-AIC averaging	0.734	0.886	0.846	0.892	$0.902^{\dagger}$	0.894
	AIC selection	0.863	0.951	0.958	0.952	1.013	0.954
	S-BIC averaging	$0.591^{\dagger}$	$0.809^{\dagger}$	$0.792^{\dagger}$	$0.822^{\dagger}$	0.929	$0.825^{\dagger}$
	BIC selection	0.659	0.877	0.931	0.885	1.113	0.876
μ_3	S-AIC averaging	0.719	0.870	0.873	0.906	$0.964^{\dagger}$	0.924
	AIC selection	0.873	0.943	0.993	0.968	1.077	0.975
	S-BIC averaging	$0.566^{\dagger}$	$0.778^{\dagger}$	$0.841^{\dagger}$	$0.847^{\dagger}$	1.038	$0.882^{\dagger}$
	BIC selection	0.643	0.859	0.999	0.915	1.255	0.933
n=200							
μ_1	S-AIC averaging	0.842	0.924	0.953	0.956	0.954	0.957
	AIC selection	0.915	0.969	1.039	0.999	1.036	0.998
	S-BIC averaging	$0.762^{\dagger}$	$0.861^{\dagger}$	$0.944^{\dagger}$	$0.925^{\dagger}$	$0.945^{\dagger}$	$0.925^{\dagger}$
	BIC selection	0.777	0.904	1.043	0.967	1.046	0.963
μ_2	S-AIC averaging	0.690	0.859	0.815	0.872	$0.912^{\dagger}$	0.885
	AIC selection	0.819	0.938	0.957	0.942	1.042	0.964
	S-BIC averaging	$0.536^{\dagger}$	$0.746^{\dagger}$	$0.782^{\dagger}$	$0.775^{\dagger}$	1.033	$0.798^{\dagger}$
	BIC selection	0.588	0.812	0.917	0.830	1.243	0.852
μ_3	S-AIC averaging	0.651	0.846	0.846	0.895	$0.974^{\dagger}$	0.916
	AIC selection	0.770	0.943	0.983	0.959	1.123	0.974
	S-BIC averaging	$0.480^{\dagger}$	$0.723^{\dagger}$	$0.834^{\dagger}$	$0.827^{\dagger}$	1.141	$0.876^{\dagger}$
	BIC selection	0.537	0.800	1.003	0.902	1.383	0.945

Table 3: Point estimates in real data analysis

	μ_1	μ_2	μ_3	μ_4
S-AIC	0.44318	0.18317	0.49939	2.42269
S-BIC	0.44331	0.18296	0.49926	2.42623
AIC (BIC)	0.44040	0.18850	0.50245	2.33634
Full Model	0.43973	0.18917	0.50261	2.32457

Table 4: 95% Confidence intervals in real data analysis

		Limits	
		Lower	Upper
μ_1			
	AIC (BIC)	0.42338	0.45742
	S-AIC/S-BIC/Full Model	0.42345	0.45601
μ_2			
	AIC (BIC)	0.16626	0.21074
	S-AIC/S-BIC/Full Model	0.16749	0.21084
μ_3			
	AIC (BIC)	0.49103	0.51386
	S-AIC/S-BIC/Full Model	0.49252	0.51270
μ_4			
	AIC (BIC)	1.99905	2.67364
	S-AIC	1.96064	2.60450
	S-BIC	2.00915	2.65301
	Full Model	2.00264	2.64650