Interval Estimation by Frequentist Model Averaging

Haiying Wang[a], Sherry Z.F. Zhou[b]

[a]*Department of Statistics, University of Missouri, Columbia, Missouri 65211, USA*
[b]*Department of Management Sciences, City University of Hong Kong, Kowloon, H.K.*

Abstract

An important contribution to the literature on frequentist model averaging (FMA) 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. They also proposed a simple method for constructing confidence intervals of the unknown parameters. Our paper shows that the confidence intervals based on the FMA estimator suggested by Hjort and Claeskens (2003) are asymptotically equivalent to that obtained from the full model under both parametric and the varying-coefficient partially linear models. Thus, as long as interval estimation rather than point estimation is concerned, the confidence interval based on the full model already fulfills the objective and model averaging provides no additional useful information.

Keywords: Model Averaging, Confidence Interval, Parametric Model, Semi-parametric Model, Asymptotic Equivalence

1. Introduction

Model selection aims to choose the best single model among all candidates that can be used to fit the given set of data. Well-known criteria developed for searching the "best" model include the AIC (Akaike, 1973), Mallows' C_p (Mallows, 1973), BIC (Schwarz, 1978), RIC (Foster and George, 1994), FIC (Claeskens and Hjort, 2003), among others. No matter which criterion is adopted, the searching process recognizes the existence of more than one plausible model structure, thus additional uncertainty is introduced associated with the choice of model. However, this uncertainty is seldom taken into account when subsequent inferences are made based on the chosen model. Hence, an overconfident inference about the unknown parameters results, such

Email addresses: hwzq7@mail.missouri.edu (Haiying Wang), zhefzhou@cityu.edu.hk (Sherry Z.F. Zhou)

Preprint submitted to Communications in Statistics, Theory and MethodsNovember 21, 2011

as underestimated variances and too optimistic confidence intervals (Hjort and Claeskens, 2003; Wan and Zou, 2003; Danilov and Magnus, 2004; Claeskens and Hjort, 2008; Liang et al., 2011; Zhang et al., 2012).

An alternative approach to model selection is model averaging, where a weighted average estimator is obtained across a set of plausible models with weights indicating the degree to which each model is trusted. Without attaching to a single model, model averaging can incorporate the uncertainty aforementioned and thus does not suffer from the distortions usually associated with model selection. In addition, model averaging avoids the possibility of selecting one very poor model, and thus holds promise for reducing the risk in estimation (Leung and Barron, 2006).

Compared to its Bayesian counterpart which has long been popular among statisticians (Hoeting et al., 1999; Clyde and George, 2004), the studies on frequentist model averaging (FMA) are mostly of recent vintage. Unlike Bayesian model averaging (BMA), FMA need not specify any prior distribution; yet how to determine the optimal weighting schemes by a data-driven approach is the biggest challenge. Since recently, a great deal of work has been done in this area and important progress has been achieved. For example, Buckland et al. (1997) suggested combining models using the exponent of the negative of the AIC value as the weight for each candidate model. Yang (2001, 2003) developed an adaptive regression mixing method, and Yuan and Yang (2005) further built on this method by implementing a model screening procedure prior to model combinations. Leung and Barron (2006) proposed a weight choice scheme based on risk minimization. Hansen (2007, 2008) and Wan et al. (2010) investigated the properties of a model average estimator based on the Mallows' criterion. Magnus et al. (2010, 2011) considered a model average estimator that is semi-Bayesian and semi-frequentist. Schomaker et al. (2010) considered FMA when observations are partially missing. Liang et al. (2011) developed a weighting mechanism for FMA estimators that exhibits optimality properties in terms of the mean squared error (MSE) of the estimators. Of particular relevance to the current study is the work of Hjort and Claeskens (2003) (hereafter referred to as HC 2003), who suggested a local misspecification framework for studying the limiting distributions and asymptotic risk properties of model selection and averaging estimators in parametric models. HC (2003)'s approach has been extended to other model frameworks such as the Cox's proportional hazards models (Hjort and Claeskens, 2006), general semi-parametric models (Claeskens and Carroll, 2007) (hereafter referred to as CC 2007), and the generalized additive partial linear models (Zhang and Liang, 2011). Useful surveys of this rapidly growing literature can be found in Claeskens and Hjort (2008) and Wang et al. (2009). A recent example of

FMA in the applied literature was given in Wan and Zhang (2009).

One major focus of the aforementioned studies was on the comparisons of risk performances between model selection estimators and model average estimators obtained from different weight choices. However, the comparisons were mainly carried out for point estimators rather than interval estimators. HC (2003) and CC (2007) investigated the interval estimation based on model averaging. The former work proposed a simple method for constructing confidence intervals for the unknown parameters, and showed that the resultant intervals have a coverage probability that converges to the intended level in large samples. CC (2007) demonstrated that all the results of HC (2003) hold in the semi-parametric context, and the simulation study revealed that the intervals based on AIC weights are remarkably close to those derived from fitting the full model.

In the current paper, we prove that the confidence intervals constructed along the lines of HC (2003) are asymptotically equivalent to those obtained from the full model. Kabaila and Leeb (2006) stated this result as an observation without giving a proof due to its simplicity. However, the work in Kabaila and Leeb (2006) was conducted in the parametric context. We prove that this asymptotic equivalence result holds not only in parametric models, but within the varying-coefficient partially linear (VCPL) models as well. The VCPL model (Zhang et al., 2002; Fan and Huang, 2005) has been an important development in the semi-parametric literature in recent years. It allows the covariates in the model to interact in a flexible way, and includes various semi-parametric models, such as the varying-coefficient model (Hastie and Tibshirani, 1993) and the partially linear model (Engle et al., 1986), as special cases. Hence, our result implies that as long as interval estimation rather than point estimation is concerned, the confidence interval based on the full model already fulfills the objective, and model averaging provides no additional useful information.

The rest of the paper is organized as follows. Section 2 describes the model setup. Section 3 presents the main theoretical results in the parametric context. Section 4 extends the analysis to the VCPL model framework. Section 5 offers our concluding remarks and the Appendix contains the proofs of the theorems.

2. Model framework and basic results

In this section, we summarize the key results of HC (2003) which provides the basis of our analysis. Assume i.i.d observations $Y_1, ..., Y_n$ come from density

f which takes the form

$$f_{\text{true}}(y) = f(y, \theta, \gamma) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}), \tag{1}$$

where θ and γ are respectively p and q dimensional vectors, θ_0 is the unknown true value of θ, γ_0 is fixed and known, and $\delta = (\delta_1, ..., \delta_q)^\top$ is a q dimensional vector representing the degree of departure from the null model $f(y, \theta, \gamma_0)$. The inclusion of θ in the model is mandatory, while that of $\gamma = (\gamma_1, ..., \gamma_q)^\top$ is optional. This is the local misspecification framework suggested in HC (2003). The parameter of interest is $\mu_{\text{true}} = \mu(\theta, \gamma) = \mu(\theta, \gamma_0 + \delta/\sqrt{n})$. Clearly, there are 2^q submodels to consider in which $\delta_j = 0$ for $j \in S^{\text{C}}$ and others are not, where S is a subset of $\{1, ..., q\}$ and S^{C} is the complement of S. The submodel S includes exactly the γ_j parameters for $j \in S$. Let $\hat{\theta}_S$ and $\hat{\gamma}_S$ be respectively the maximum likelihood estimators (MLEs) of θ and γ_js in model S. The MLE of $\mu = \mu(\theta, \gamma)$ is thus $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^{\text{C}}})$, where $\hat{\gamma}_j$ for $j \in S$ are included while other elements of γ are kept at their null values of γ_0.

Denote J_{full} as the $(p+q) \times (p+q)$ information matrix of the full model evaluated at the null point (θ_0, γ_0), then

$$J_{\text{full}} = Var_0 \begin{pmatrix} U(Y) \\ V(Y) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$$

with inverse

$$J_{\text{full}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

where $Var_0(M)$ is the variance matrix of M evaluated at the null point, and $U(y) = \partial \log f(y, \theta_0, \gamma_0)/\partial\theta$ and $V(y) = \partial \log f(y, \theta_0, \gamma_0)/\partial\gamma$ are the p and q dimensional score functions, respectively.

Let π_S be the projection matrix such that $\pi_S v = v_S$; that is, the vector $v = (v_1, ..., v_q)^\top$ is mapped to its subvector v_S with components v_j, $j \in S$. Define $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, $K_S = (\pi_S K^{-1} \pi_S^\top)^{-1}$, $H_S = K^{-1/2}\pi_S^\top K_S \pi_S K^{-1/2}$, and $\omega = J_{10}J_{00}^{-1}\partial\mu/\partial\theta - \partial\mu/\partial\gamma$. Then from HC (2003) (Lemmas 3.2 and 3.3), we have

$$D_n = \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D \sim N_q(\delta, K), \tag{2}$$

and

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S \equiv \left(\frac{\partial\mu}{\partial\theta}\right)^\top J_{00}^{-1} M + \omega^\top(\delta - K^{1/2} H_S K^{-1/2} D), \tag{3}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\hat{\gamma}_{\text{full}}$ is the MLE of γ under the full model, $M \sim N_p(0, J_{00})$ is independent of D, and the partial derivatives are evaluated at the null point (θ_0, γ_0).

With each submodel estimator being a candidate, the model averaging estimator takes the form

$$\hat{\mu} = \sum_S c(S \mid D_n)\hat{\mu}_S, \tag{4}$$

where $c(S \mid D_n)$'s are weight functions and it is required that $\sum_S c(S|d) = 1$ for each d. Theorem 4.1 in HC (2003) showed that

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \xrightarrow{d} \Lambda \equiv \left(\frac{\partial \mu}{\partial \theta}\right)^\top J_{00}^{-1} M + \omega^\top \{\delta - \hat{\delta}(D)\}, \tag{5}$$

where $\hat{\delta}(D) = K^{1/2}\{\sum_S c(S \mid D) H_S\} K^{-1/2} D$. Moreover, the asymptotic risk of $\hat{\mu}$ is

$$R_a(\hat{\mu}, \mu) = E(\Lambda^2) = \tau_0^2 + E\{\omega^\top \hat{\delta}(D) - \omega^\top \delta\}^2,$$

where $\tau_0^2 = (\partial \mu / \partial \theta)^\top J_{00}^{-1} (\partial \mu / \partial \theta)$.

The asymptotic distribution of the model averaging estimator $\hat{\mu}$ is non-normal. HC (2003) then suggested an approach of constructing the confidence interval based on the model averaging estimator. The confidence limits are developed as

$$\begin{aligned} \text{low}_n &= \hat{\mu} - \hat{\omega}^\top [D_n - \hat{\delta}(D_n)]/\sqrt{n} - u\hat{\kappa}/\sqrt{n}, \\ \text{up}_n &= \hat{\mu} - \hat{\omega}^\top [D_n - \hat{\delta}(D_n)]/\sqrt{n} + u\hat{\kappa}/\sqrt{n}, \end{aligned} \tag{6}$$

where $\hat{\omega}$ and $\hat{\kappa}$ are consistent estimators of ω and $\kappa = (\tau_0^2 + \omega^\top K \omega)^{1/2}$, respectively, and u is a standard normal quantile. They showed that the coverage probability of this proposed interval converges to the intended level in large samples, i.e.,

$$\Pr\big\{\mu_{\text{true}} \in (\text{low}_n, \text{up}_n)\big\} \to 1 - \alpha,$$

where $1 - \alpha$ is the nominal coverage probability.

3. Asymptotic equivalence of confidence intervals

This section shows that the confidence interval with lower and upper limits (6) is asymptotically equivalent to that constructed based on the full model estimator that asymptotically follows a normal distribution. It is obtained from (3) that the limiting variable corresponding to the full model is $\Lambda_{\text{full}} \equiv$

$\left(\frac{\partial\mu}{\partial\theta}\right)^\top J_{00}^{-1}M + \omega^\top(\delta - D)$ which has an $N(0, \kappa^2)$ distribution. Accordingly, the lower and upper limits of the confidence interval obtained from the full model estimation are respectively

$$\begin{aligned} \text{low}_{\text{full}} &= \hat{\mu}_{\text{full}} - u\hat{\kappa}/\sqrt{n}, \\ \text{up}_{\text{full}} &= \hat{\mu}_{\text{full}} + u\hat{\kappa}/\sqrt{n}, \end{aligned} \tag{7}$$

where $\hat{\mu}_{\text{full}}$ denotes the estimator of μ under the full model. Note that the difference between the lengths of the intervals (6) and (7) is of order $o_P(\frac{1}{\sqrt{n}})$, and $\hat{\omega}$ and $\hat{\kappa}$ are consistent estimators. In what follows, we show that the difference between the centers of the two intervals is also of order $o_P(\frac{1}{\sqrt{n}})$. This result, once proved, indicates that the intervals (6) and (7) are asymptotically equivalent.

From HC (2003), we have

$$\begin{aligned} \begin{pmatrix} \hat{\theta}_S - \theta_0 \\ \hat{\gamma}_S - \gamma_{0,S} \end{pmatrix} &= J_S^{-1} \begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} + o_P(\frac{1}{\sqrt{n}}) \\ &= J_S^{-1} \begin{pmatrix} I & 0 \\ 0 & \pi_S \end{pmatrix} \begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix} + o_P(\frac{1}{\sqrt{n}}), \end{aligned} \tag{8}$$

where $\bar{U}_n = \frac{1}{n}\sum_{i=1}^n U(Y_i)$, $\bar{V}_n = \frac{1}{n}\sum_{i=1}^n V(Y_i)$, and

$$J_S = Var_0 \left(\begin{array}{c} U(Y) \\ V_S(Y) \end{array} \right) = \left(\begin{array}{cc} J_{00} & J_{01}\pi_S^\top \\ \pi_S J_{10} & \pi_S J_{11} \pi_S^\top \end{array} \right).$$

In particular,

$$\begin{pmatrix} \hat{\theta}_{\text{full}} - \theta_0 \\ \hat{\gamma}_{\text{full}} - \gamma_0 \end{pmatrix} = J_{\text{full}}^{-1} \begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix} + o_P(\frac{1}{\sqrt{n}}), \tag{9}$$

where $\hat{\theta}_{\text{full}}$ is the MLE of θ under the full model.

From equations (8) and (9), we obtain that

$$\begin{aligned}\begin{pmatrix}\hat\theta_S-\theta_0\\ \hat\gamma_S-\gamma_{0,S}\end{pmatrix} =& J_S^{-1}\begin{pmatrix}I & 0\\ 0 & \pi_S\end{pmatrix} J_{\text{full}}\begin{pmatrix}\hat\theta_{\text{full}}-\theta_0\\ \hat\gamma_{\text{full}}-\gamma_0\end{pmatrix}+o_P(\frac{1}{\sqrt{n}})\\ =&\begin{pmatrix}J_{00}^{-1}+J_{00}^{-1}J_{01}\pi_S^\top K_S\pi_S J_{10}J_{00}^{-1} & -J_{00}^{-1}J_{01}\pi_S^\top K_S\\ -K_S\pi_S J_{10}J_{00}^{-1} & K_S\end{pmatrix}\times\\ &\begin{pmatrix}I & 0\\ 0 & \pi_S\end{pmatrix}\begin{pmatrix}J_{00} & J_{01}\\ J_{10} & J_{11}\end{pmatrix}\begin{pmatrix}\hat\theta_{\text{full}}-\theta_0\\ \hat\gamma_{\text{full}}-\gamma_0\end{pmatrix}+o_P(\frac{1}{\sqrt{n}})\\ =&\begin{pmatrix}I & J_{00}^{-1}J_{01}(I-\pi_S^\top K_S\pi_S K^{-1})\\ 0 & K_S\pi_S K^{-1}\end{pmatrix}\begin{pmatrix}\hat\theta_{\text{full}}-\theta_0\\ \hat\gamma_{\text{full}}-\gamma_0\end{pmatrix}+o_P(\frac{1}{\sqrt{n}}).\end{aligned}$$

Therefore, we have

$$\hat\theta_S-\theta_0=(\hat\theta_{\text{full}}-\theta_0)+J_{00}^{-1}J_{01}(I-\pi_S^\top K_S\pi_S K^{-1})(\hat\gamma_{\text{full}}-\gamma_0)+o_P(\frac{1}{\sqrt{n}}), \tag{10}$$

and

$$\hat\gamma_S-\gamma_{0,S}=K_S\pi_S K^{-1}(\hat\gamma_{\text{full}}-\gamma_0)+o_P(\frac{1}{\sqrt{n}}). \tag{11}$$

By a Taylor series expansion,

$$\begin{aligned}\hat\mu_S&=\mu(\hat\theta_S,\hat\gamma_S)\\ &=\mu(\theta_0,\gamma_0)+\Big(\frac{\partial\mu}{\partial\theta}\Big)^\top(\hat\theta_S-\theta_0)+\Big(\frac{\partial\mu}{\partial\gamma}\Big)^\top\pi_S^\top(\hat\gamma_S-\gamma_{0,S})+o_P(\frac{1}{\sqrt{n}}),\end{aligned} \tag{12}$$

and

$$\begin{aligned}\hat\mu_{\text{full}}&=\mu(\hat\theta_{\text{full}},\hat\gamma_{\text{full}})\\ &=\mu(\theta_0,\gamma_0)+\Big(\frac{\partial\mu}{\partial\theta}\Big)^\top(\hat\theta_{\text{full}}-\theta_0)+\Big(\frac{\partial\mu}{\partial\gamma}\Big)^\top(\hat\gamma_{\text{full}}-\gamma_0)+o_P(\frac{1}{\sqrt{n}}).\end{aligned} \tag{13}$$

Substituting equations (10) - (13) into (4), we derive that

$$
\begin{aligned}
\hat{\mu} =&\mu(\theta_0,\gamma_0) + \Big(\frac{\partial \mu}{\partial \theta}\Big)^\top \sum_S c(S|D_n)(\hat{\theta}_S - \theta_0) + \Big(\frac{\partial \mu}{\partial \gamma}\Big)^\top \sum_S c(S|D_n)\pi_S^\top(\hat{\gamma}_S - \gamma_{0,S}) + o_P(\frac{1}{\sqrt{n}}) \\
=&\mu(\theta_0,\gamma_0) + \Big(\frac{\partial \mu}{\partial \theta}\Big)^\top \Big[(\hat{\theta}_{\text{full}} - \theta_0) + J_{00}^{-1}J_{01}\big(I - \sum_S c(S|D_n)\pi_S^\top K_S \pi_S K^{-1}\big)(\hat{\gamma}_{\text{full}} - \gamma_0)\Big] \\
&+ \Big(\frac{\partial \mu}{\partial \gamma}\Big)^\top \Big[\sum_S c(S|D_n)\pi_S^\top K_S \pi_S K^{-1}(\hat{\gamma}_{\text{full}} - \gamma_0)\Big] + o_P(\frac{1}{\sqrt{n}}) \\
=&\mu(\theta_0,\gamma_0) + \Big(\frac{\partial \mu}{\partial \theta}\Big)^\top (\hat{\theta}_{\text{full}} - \theta_0) + \Big(\frac{\partial \mu}{\partial \gamma}\Big)^\top (\hat{\gamma}_{\text{full}} - \gamma_0) \\
&+ \Big(\frac{\partial \mu}{\partial \theta}\Big)^\top J_{00}^{-1}J_{01}(\hat{\gamma}_{\text{full}} - \gamma_0) - \Big(\frac{\partial \mu}{\partial \gamma}\Big)^\top (\hat{\gamma}_{\text{full}} - \gamma_0) \\
&- \Big[\Big(\frac{\partial \mu}{\partial \theta}\Big)^\top J_{00}^{-1}J_{01} - \Big(\frac{\partial \mu}{\partial \gamma}\Big)^\top\Big] \sum_S c(S|D_n)\pi_S^\top K_S \pi_S K^{-1}(\hat{\gamma}_{\text{full}} - \gamma_0) + o_P(\frac{1}{\sqrt{n}}) \\
=&\hat{\mu}_{\text{full}} + \omega^\top \left[I - K^{1/2}\left\{\sum_S c(S|D_n)K^{-1/2}\pi_S^\top K_S \pi_S K^{-1/2}\right\}K^{-1/2}\right](\hat{\gamma}_{\text{full}} - \gamma_0) + o_P(\frac{1}{\sqrt{n}}) \\
=&\hat{\mu}_{\text{full}} + \omega^\top[D_n - \hat{\delta}(D_n)]/\sqrt{n} + o_P(\frac{1}{\sqrt{n}}).
\end{aligned} \tag{14}
$$

Upon the comparison among equations (6), (7) and (14), we obtain that $\text{low}_n = \text{low}_{\text{full}} + o_P(\frac{1}{\sqrt{n}})$ and $\text{up}_n = \text{up}_{\text{full}} + o_P(\frac{1}{\sqrt{n}})$. Thus, the two confidence intervals, based on the model averaging estimator and the full model estimator, are asymptotically equivalent. This result is obtained in the parametric context; we extend this analysis to the semi-parametric model framework in the next section.

4. Extension to varying-coefficient partially linear models

Consider varying-coefficient partially linear (VCPL) model (Zhang et al., 2002; Fan and Huang, 2005):

$$Y = Z^\top \beta + X^\top \alpha(T) + \varepsilon, \tag{15}$$

where Y is the response variable and (Z, X, T) are covariates, $\beta = (\tilde{\theta}^\top, \tilde{\gamma}^\top)^\top$ is a $(p+q)$ dimensional parametric coefficient vector with $\tilde{\theta}$ and $\tilde{\gamma}$ being p and q dimensional respectively, $\alpha(\cdot) = (\alpha_1(\cdot), ..., \alpha_r(\cdot))^\top$ is an r dimensional unknown coefficient function, and ε is a random error vector with mean 0 and variance σ^2, and it is independent of (Z, X, T). Following Fan and Huang (2005), we assume that the dimension of T is one. The VCPL model includes many common models as special cases. For example, when $\beta \equiv 0$, it reduces to the varying-coefficient model (Hastie and Tibshirani, 1993), and when $r = 1$ and $X \equiv 1$, it becomes the partially linear model (Engle et al., 1986).

Similar to the local mis-specification framework in the preceding analysis under parametric setup, we assume $\beta = \begin{pmatrix} \tilde{\theta} \\ \tilde{\gamma} \end{pmatrix} = \begin{pmatrix} \tilde{\theta} \\ \tilde{\delta}/\sqrt{n} \end{pmatrix}$, where the parameters $\tilde{\delta}_1, ..., \tilde{\delta}_q$ represent the degrees of model departure in directions $1, ..., q$ from the narrow model for which $\tilde{\gamma} = 0$ as in CC (2007).

For any submodel S, it includes all elements of $\tilde{\theta}$, while contains only certain elements of $\tilde{\gamma}$. Let $\tilde{\theta}_S$ and $\tilde{\gamma}_S$ be the coefficients in model S. The profile least-squares method suggested in Fan and Huang (2005) can be used to estimate $\beta_S = (\tilde{\theta}_S^\top, \tilde{\gamma}_S^\top)^\top$ and the estimator is denoted by $\hat{\beta}_S = (\hat{\tilde{\theta}}_S^\top, \hat{\tilde{\gamma}}_S^\top)^\top$. Define $M_{n1} = \sqrt{n}(\hat{\tilde{\theta}}_{\text{full}} - \theta_0)$ and $M_{n2} = \sqrt{n}\hat{\tilde{\gamma}}_{\text{full}}$, where $\tilde{\theta}_{\text{full}}$ and $\tilde{\gamma}_{\text{full}}$ are the coefficients in the full model. Let $B = E(ZZ^\top) - E\big[E(ZX^\top|T)E(XX^\top|T)^{-1}E(XZ^\top|T)\big]$ and partition B comfortably with the dimensions of $\tilde{\theta}$ and $\tilde{\gamma}$ into $\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$. A consistent estimator of B is $B_n = \begin{pmatrix} B_{n11} & B_{n12} \\ B_{n21} & B_{n22} \end{pmatrix} = \mathbf{Z}^\top(I_n - \mathbf{S})^\top(I_n - \mathbf{S})\mathbf{Z}$, where $\mathbf{Z} = (Z_1, ..., Z_n)^\top$ is the sample matrix of Z, and $\mathbf{S}$ is the smoothing matrix as defined in Fan and Huang (2005). Note that $M_1 + B_{11}^{-1}B_{12}(M_2 - \delta)$ and M_2 are stochastically independent.

Consider the parameter of interest $\tilde{\mu}_{\text{true}} = \tilde{\mu}(\tilde{\theta}, \tilde{\gamma})$. Let the estimator based on the reduced model S be $\hat{\tilde{\mu}}_S = \tilde{\mu}(\hat{\tilde{\theta}}_S, \hat{\tilde{\gamma}}_S)$. The following theorem is then obtained.

Theorem 1. *Assume that $\tilde{\mu}$ is differentiable at $\beta_0 = \begin{pmatrix} \tilde{\theta}_0 \\ 0_{q\times 1} \end{pmatrix}$. If conditions (C1) - (C6) in the Appendix are satisfied, and ε and (Z, X, T) are independent, then we have*

$$\begin{aligned}\sqrt{n}(\hat{\tilde{\mu}}_S - \tilde{\mu}_{\text{true}}) \xrightarrow{d} \tilde{\Lambda}_S =& \tilde{\mu}_{\tilde{\theta}}^\top [M_1 + B_{11}^{-1}B_{12}(M_2 - \tilde{\delta})] \\ &+ \tilde{\omega}^\top\left(\tilde{\delta} - A^{-1/2}\tilde{H}_s A^{1/2} M_2\right),\end{aligned} \tag{16}$$

where M_1 and M_2 are the limiting variables of M_{n1} and M_{n2} respectively, $\tilde{\mu}_{\tilde{\theta}} = \frac{\partial\mu_{(\tilde{\theta}_0,0)}}{\partial\tilde{\theta}}$, $\tilde{\mu}_{\tilde{\gamma}} = \frac{\partial\tilde{\mu}_{(\tilde{\theta}_0,0)}}{\partial\tilde{\gamma}}$, $\tilde{\omega} = B_{21}B_{11}^{-1}\tilde{\mu}_{\tilde{\theta}} - \tilde{\mu}_{\tilde{\gamma}}$, $A = B_{22} - B_{21}B_{11}^{-1}B_{12}$, and $\tilde{H}_S = A^{1/2}\pi_S^\top(\pi_S A \pi_S^\top)^{-1}\pi_S A^{1/2}$.

Consider the FMA estimator

$$\hat{\tilde{\mu}} = \sum_S \tilde{c}(S|M_{n2})\hat{\tilde{\mu}}_S, \tag{17}$$

where $\tilde{c}(S|M_{n2})$'s are weight functions and it is required that $\sum_S \tilde{c}(S|d) = 1$ for each d. The asymptotic properties of the estimator $\hat{\tilde{\mu}}$ are illustrated in the following theorem.

Theorem 2. *Assume that $\tilde{\mu}$ is differentiable at $\tilde{\beta}_0$, and $c(S|\cdot)$ is a.s. continuous. If conditions (C1) - (C6) in the Appendix hold, and ε and (Z, X, T) are independent, then we have*

$$\begin{aligned}\sqrt{n}(\hat{\tilde{\mu}} - \tilde{\mu}_{\text{true}}) \xrightarrow{d} \tilde{\Lambda} &= \tilde{\mu}_{\tilde{\theta}}^\top \Big[M_1 + B_{11}^{-1} B_{12}(M_2 - \tilde{\delta})\Big] + \tilde{\omega}^\top \Big(\tilde{\delta} - Q(M_2)M_2\Big),\\ \mathbf{E}(\tilde{\Lambda}) &= \tilde{\omega}^\top \Big(\tilde{\delta} - \mathbf{E}[Q(M_2)M_2]\Big), \textit{and}\\ \mathbf{Var}(\tilde{\Lambda}) &= \tilde{\tau}_0^2 + \tilde{\omega}^\top \mathbf{Var}\big[Q(M_2)M_2\big]\tilde{\omega},\end{aligned}$$

where $\tilde{\tau}_0^2 = \tilde{\mu}_{\tilde{\theta}}^\top B_{11}^{-1} \tilde{\mu}_{\tilde{\theta}}$ and $Q(M_2) = A^{-1/2}\big(\sum_S \tilde{c}(S|M_2)\tilde{H}_S\big)A^{1/2}$.

This theorem reveals that the asymptotic distribution of the model averaging estimator is non-normal; we thus follow the approach suggested in HC (2003) and CC (2007) and develop the confidence limits for the interval estimate of $\tilde{\mu}$ as

$$\begin{aligned}\widetilde{\text{low}}_n &= \hat{\tilde{\mu}} - \hat{\tilde{\omega}}^\top [M_{n2} - Q(M_{n2})M_{n2}]/\sqrt{n} - u\hat{\tilde{\kappa}}/\sqrt{n},\\ \widetilde{\text{up}}_n &= \hat{\tilde{\mu}} - \hat{\tilde{\omega}}^\top [M_{n2} - Q(M_{n2})M_{n2}]/\sqrt{n} + u\hat{\tilde{\kappa}}/\sqrt{n},\end{aligned} \tag{18}$$

where $\hat{\tilde{\omega}}$ and $\hat{\tilde{\kappa}}$ are consistent estimators of $\tilde{\omega}$ and $\tilde{\kappa} = \sqrt{(\tilde{\mu}_{\tilde{\theta}}^\top, \tilde{\mu}_{\tilde{\gamma}}^\top)B^{-1}(\tilde{\mu}_{\tilde{\theta}}^\top, \tilde{\mu}_{\tilde{\gamma}}^\top)^\top}$, respectively, $Q_n(M_{n2}) = A_n^{-1/2}\big(\sum_S c(S|M_{n2})\tilde{H}_{nS}\big)A_n^{1/2}$, $A_n = B_{n22} - B_{n21}B_{n11}^{-1}B_{n12}$, and $\tilde{H}_{nS} = A_n^{1/2}\pi_S^\top(\pi_S A_n \pi_S^\top)^{-1}\pi_S A_n^{1/2}$. The confidence interval bounded by $\widetilde{\text{low}}_n$ and $\widetilde{\text{high}}_n$ covers the true parameter with probability $\Pr\big\{\tilde{\mu}_{true} \in \{\widetilde{\text{low}}_n, \widetilde{\text{high}}_n\}\big\}$ $= Pr\{-u \leq T_n \leq u\}$, where

$$T_n = \frac{\sqrt{n}(\hat{\tilde{\mu}} - \tilde{\mu}_{true}) - \hat{\tilde{\omega}}^\top [M_{n2} - Q_n(M_{n2})M_{n2}]}{\hat{\tilde{\kappa}}}.$$

From the Continuous mapping theorem and Slutsky theorem, we have

$$\{\sqrt{n}(\hat{\mu} - \mu_{true}), M_{n2}\} \xrightarrow{d} \Big\{\tilde{\mu}_{\tilde{\theta}}^\top \Big[M_1 + B_{11}^{-1}B_{12}(M_2 - \tilde{\delta})\Big] + \tilde{\omega}^\top \Big(\tilde{\delta} - Q(M_2)M_2\Big), M_2\Big\}.$$

Thus,

$$T_n \xrightarrow{d} \frac{\tilde{\mu}_{\tilde{\theta}}^\top \Big[M_1 + B_{11}^{-1}B_{12}(M_2 - \tilde{\delta})\Big] + \tilde{\omega}^\top(\tilde{\delta} - M_2)}{\tilde{\kappa}} = \frac{\tilde{\mu}_{\tilde{\theta}}^\top M_1 + \tilde{\mu}_{\tilde{\gamma}}^\top (M_2 - \tilde{\delta})}{\tilde{\kappa}}.$$

Since $\tilde{\mu}_{\tilde{\theta}}^\top M_1 + \tilde{\mu}_{\tilde{\gamma}}^\top (M_2 - \tilde{\delta}) \sim N(0, \tilde{\kappa}^2)$, we have $Pr\{-u \leq T_n \leq u\} \to 1 - \alpha$.

Denoting $\hat{\tilde{\mu}}_{\text{full}}$ as the estimator of $\tilde{\mu}$ under the full model, then we obtain from Theorem 1 that

$$\begin{aligned}\sqrt{n}(\hat{\tilde{\mu}}_{\text{full}} - \tilde{\mu}_{true}) \xrightarrow{d} & \tilde{\mu}_{\tilde{\theta}}^{\top}[M_1 + B_{11}^{-1}B_{12}(M_2 - \tilde{\delta})] + \tilde{\omega}^{\top}\Big(\tilde{\delta} - M_2\Big) \\ &= \tilde{\mu}_{\tilde{\theta}}^{\top} M_1 + \tilde{\mu}_{\tilde{\gamma}}^{\top}(M_2 - \tilde{\delta}).\end{aligned}$$

The confidence limits of $\tilde{\mu}_{true}$ based on $\hat{\mu}_{\text{full}}$ are thus

$$\begin{cases}\widetilde{\text{low}}_{\text{full}} = \hat{\tilde{\mu}}_{\text{full}} - u\hat{\tilde{\kappa}}/\sqrt{n} \\ \widetilde{\text{up}}_{\text{full}} = \hat{\tilde{\mu}}_{\text{full}} + u\hat{\tilde{\kappa}}/\sqrt{n}.\end{cases} \tag{19}$$

Through a Taylor series expansion, we have

$$\hat{\tilde{\mu}}_S = \tilde{\mu}(\hat{\tilde{\theta}}_S, \hat{\tilde{\gamma}}_S) = \tilde{\mu}(\tilde{\theta}_0, 0) + \begin{pmatrix}\tilde{\mu}_{\tilde{\theta}} \\ \pi\tilde{\mu}_{\tilde{\gamma}}\end{pmatrix}^{\top} \begin{pmatrix}\hat{\tilde{\theta}}_S - \tilde{\theta}_0 \\ \hat{\tilde{\gamma}}_S\end{pmatrix} + o_P(\frac{1}{\sqrt{n}}). \tag{20}$$

Substituting equation (20) into the definition of $\hat{\tilde{\mu}}$ in equation (17), we then obtain that

$$\begin{aligned}\hat{\tilde{\mu}} &= \tilde{\mu}(\tilde{\theta}_0, 0) + \sum_S \tilde{c}(S|M_{n2}) \begin{pmatrix}\tilde{\mu}_{\tilde{\theta}} \\ \pi\tilde{\mu}_{\tilde{\gamma}}\end{pmatrix}^{\top} \begin{pmatrix}\hat{\tilde{\theta}}_S - \tilde{\theta}_0 \\ \hat{\tilde{\gamma}}_S\end{pmatrix} + o_P(\frac{1}{\sqrt{n}}) \\ &= \tilde{\mu}(\tilde{\theta}_0, 0) + \tilde{\mu}_{\tilde{\theta}}^{\top}(\hat{\tilde{\theta}}_{\text{full}} - \tilde{\theta}_0) + \sum_S \tilde{c}(S|M_{n2})\Big(\tilde{\mu}_{\tilde{\theta}}^{\top} C_{ns}\hat{\tilde{\gamma}}_{\text{full}} + \tilde{\mu}_{\tilde{\gamma}}^{\top}\pi^{\top}\hat{\tilde{\gamma}}_S\Big) + o_P(\frac{1}{\sqrt{n}}) \\ &= \tilde{\mu}(\tilde{\theta}_0, 0) + \tilde{\mu}_{\tilde{\theta}}^{\top}(\hat{\tilde{\theta}}_{\text{full}} - \tilde{\theta}_0) + \sum_S \tilde{c}(S|M_{n2})\Big(\tilde{\mu}_{\tilde{\theta}}^{\top} C_{ns}\hat{\tilde{\gamma}}_{\text{full}} + \tilde{\mu}_{\tilde{\gamma}}^{\top} A_n^{-1/2}\tilde{H}_{nS}A_n^{1/2}\hat{\tilde{\gamma}}_{\text{full}}\Big) + o_P(\frac{1}{\sqrt{n}}) \\ &= \tilde{\mu}(\tilde{\theta}_0, 0) + \tilde{\mu}_{\tilde{\theta}}^{\top}(\hat{\tilde{\theta}}_{\text{full}} - \tilde{\theta}_0) \\ &\quad + \sum_S \tilde{c}(S|M_{n2})\Big\{\tilde{\mu}_{\tilde{\theta}}^{\top} B_{11}^{-1}B_{12}(I_q - A_n^{-1/2}\tilde{H}_{nS}A_n^{1/2}) + \tilde{\mu}_{\tilde{\gamma}}^{\top} A_n^{-1/2}\tilde{H}_{nS}A_n^{1/2}\Big\}\frac{M_{n2}}{\sqrt{n}} + o_P(\frac{1}{\sqrt{n}}) \\ &= \tilde{\mu}(\tilde{\theta}_0, 0) + \tilde{\mu}_{\tilde{\theta}}^{\top}(\hat{\tilde{\theta}} - \tilde{\theta}_0) + \tilde{\mu}_{\tilde{\gamma}}^{\top}\hat{\tilde{\gamma}} + \tilde{\omega}^{\top}\{M_{n2} - Q(M_{n2})M_{n2}\}/\sqrt{n} + o_P(\frac{1}{\sqrt{n}}),\end{aligned} \tag{21}$$

where $C_{ns} = B_{n11}^{-1}B_{n12}(I - A_n^{-1/2}H_{ns}A_n^{1/2})$.
By a Taylor series expansion,

$$\hat{\tilde{\mu}}_{\text{full}} = \tilde{\mu}(\tilde{\theta}_0, 0) + \tilde{\mu}_{\tilde{\theta}}^{\top}(\hat{\tilde{\theta}}_{\text{full}} - \tilde{\theta}_0) + \tilde{\mu}_{\tilde{\gamma}}^{\top}\hat{\tilde{\gamma}}_{\text{full}} + o_P(\frac{1}{\sqrt{n}}). \tag{22}$$

Therefore, we have

$$\hat{\tilde{\mu}}_{\text{full}} = \hat{\tilde{\mu}} - \tilde{\omega}^{\top}[M_{n2} - Q(M_{n2})M_{n2}]/\sqrt{n} + o_P(\frac{1}{\sqrt{n}}). \tag{23}$$

Combining equations (18), (19) and (23), we obtain that $\widetilde{\text{low}}_n = \widetilde{\text{low}}_{\text{full}} + o_P(\frac{1}{\sqrt{n}})$ and $\widetilde{\text{up}}_n = \widetilde{\text{up}}_{\text{full}} + o_P(\frac{1}{\sqrt{n}})$. Hence, the confidence interval based on the model averaging estimator is asymptotically equivalent to that based on the full model estimator.

Note that if $\tilde{\mu}$ is a linear combination of $\tilde{\theta}$ and $\tilde{\gamma}$, then the remainder in (22) vanishes. Moreover, $\tilde{\kappa}$ and $\tilde{\omega}$ are quantities associated with the full model only, thus the estimators $\hat{\tilde{\kappa}}$ and $\hat{\tilde{\omega}}$ are the same for both the full model and the model averaging. Hence, if the parameter of interest is a linear combination of regression coefficients, the confidence interval constructed from model averaging will be exactly equivalent to that obtained under the full model. This indicates that if the main concern of the investigator is interval estimation rather than point estimation, the confidence interval based on the full model is recommended. It not only serves the purpose of estimation, but provides simplicity in computation as well.

Now we examine the finite sample performance of the FMA estimator through a simulation example. The application of our method requires a proper bandwidth h for the nonparametric component in the model, because h controls the trade-off between the goodness-of-fit and the prediction capability of the nonparametric part. In particular, smaller h results in a better goodness-of-fit, while larger h increases the prediction capability. However, our main concern is the estimation of parameters in the linear component of the model, which is insensitive to the choice of the value of h (Fan and Huang, 2005). In our simulations we use a cross-validation method to choose the bandwidth, and the values of h are selected to be 0.25 and 0.15 for $n = 100$ and $n = 200$ respectively. Our simulation study is implemented on the VCPL model

$$\begin{aligned} Y &= Z^\top \tilde{\beta} + X^\top \alpha(T) + \varepsilon \\ &= Z_1\tilde{\theta}_1 + Z_2\tilde{\theta}_2 + Z_3\tilde{\theta}_3 + Z_4\tilde{\gamma}_1 + Z_5\tilde{\gamma}_2 + Z_6\tilde{\gamma}_3 + Z_7\tilde{\gamma}_4 + Z_8\tilde{\gamma}_5 \\ &\quad + X_1\sin(2\pi T) + X_2\sin(6\pi T) + \varepsilon, \end{aligned}$$

where $Z = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8)^\top$, $X = (X_1, X_2)^\top$ $\beta = (\tilde{\theta}^\top, \tilde{\gamma}^\top)^\top = ((3, 1.5, 2), \tilde{\delta}/\sqrt{n})^\top$, $\tilde{\delta} = (0, 1.5, 0, 1, 0)^\top$, Z_1,....,Z_8, X_1 and X_2 are covariates each having a standard normal distribution with ρ_{ij} being the correlation coefficients between covariates i and j of $(Z^\top, X^\top)^\top$, $i, j = 1, ..., 10$, and both T and ε follow the standard normal distribution. Three cases are considered for the correlation ρ_{ij}:
Case 1. $\rho_{ij} = 0$;
Case 2. $\rho_{ij} = 0.5^{|i-j|}$; and
Case 3. $\rho_{ij} = 0.6$.
Our parameters of interest are $\mu_{linear} = \ell^\top \theta$ with $\ell = (1, 2, 1.5, 1, 1, 1, 1, 1)^\top$ and $\mu_{ratio} = \tilde{\theta}_1/\tilde{\theta}_2$. In each case 1000 independent samples of size n are drawn.

Following Buckland et al. (1997)'s suggestion, we construct FMA estimators by using smoothed AIC (S-AIC) and smoothed BIC (S-BIC) values, defined as $\frac{\exp(-\frac{1}{2}\mathrm{AIC}_{nS})}{\sum_S \exp(-\frac{1}{2}\mathrm{AIC}_{nS})}$ and $\frac{\exp(-\frac{1}{2}\mathrm{BIC}_{nS})}{\sum_S \exp(-\frac{1}{2}\mathrm{BIC}_{nS})}$ respectively, as weights for each candidate model S. We also consider the post-model selection estimators using AIC and BIC, which are denoted as P-AIC and P-BIC estimators respectively. They can be viewed as a model averaging estimator in the form of equation (17) with indicator functions as its weights. For example, assuming that there are no ties among the AIC values, we may write P-AIC estimator as

$$\hat{\hat{\mu}}_{\mathrm{AIC}} = \sum_S I_{\{\mathrm{AIC}_S \text{ is the smallest}\}} \hat{\hat{\mu}}_S \equiv \sum_S c(S|M_{n2}) \hat{\hat{\mu}}_S,$$

where AIC_S denotes the AIC value for model S. The same result holds for the P-BIC estimator.

Table 1: Simulation results for μ_{linear}

μ_{linear}		n=100(h=0.25)			n=200(h=0.15)		
	ρ	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
MSE							
	S-AIC	0.1504	0.0749	0.1468	0.0649	0.0292	0.0538
	S-BIC	0.1542	0.0811	0.1496	0.0660	0.0311	0.0542
	P-AIC	0.1811	0.0881	0.1533	0.0688	0.0310	0.0539
	P-BIC	0.1679	0.0844	0.1528	0.0694	0.0310	0.0537
	Full model	0.1945	0.0912	0.1565	0.0721	0.0320	0.0545
C.P.							
	Full model	93.8%	93.4%	93.3%	95.3%	95.3%	95.5%
Upper							
	Full model	10.1266	9.8320	10.0018	9.6995	9.5306	9.6404
Lower							
	Full model	8.4458	8.7077	8.5418	8.6401	8.8161	8.7095

We compare the coverage probabilities (C.P.) and the upper and lower limits of confidence intervals constructed based on model averaging, post-model selection, as well as the full model. Meanwhile, we also evaluate the MSE performances of FMA estimator, post-model selection estimator, and the estimator derived from the full model. The results are reported in Tables 1 and 2. We observe that model averaging has an advantage over model selection estimators and full model estimation with respect to the MSE values, which indicates that the selection of appropriate weights is necessary for point estimates. In the case where the parameter of interest is μ_{linear} (Table 1), the confidence intervals developed from model averaging are exactly the same as those obtained from the full model, hence only the results derived under the full model are reported. When we focus on the parameter μ_{ratio}, Table 2 shows that the both the coverage probabilities and confidence limits of the intervals developed based on model averaging are very close to those obtained under the full model, and the differences in the confidence limits decrease with the sample size.

Table 2: Simulation results for μ_{ratio}

μ_{ratio}		n=100(h=0.25)			n=200(h=0.15)		
	ρ	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
MSE							
	S-AIC	0.0340	0.0668	0.0801	0.0137	0.0275	0.0327
	S-BIC	0.0339	0.0688	0.0853	0.0137	0.0280	0.0342
	P-AIC	0.0360	0.0733	0.0901	0.0140	0.0284	0.0342
	P-BIC	0.0363	0.0709	0.0845	0.0140	0.0282	0.0330
	Full model	0.0362	0.0737	0.0941	0.0139	0.0285	0.0350
C.P.							
	S-AIC	93.6%	94.1%	93.1%	95%	94.6%	95.2%
	S-BIC	93.6%	94%	93.1%	95%	94.6%	95.2%
	P-AIC	93.6%	94.1%	93.2%	95%	94.6%	95.3%
	P-BIC	93.6%	94.1%	93.2%	95%	94.6%	95.3%
	Full model	93.6%	94%	93.2%	95%	94.6%	95.3%
Upper							
	S-AIC	2.3689	2.5319	2.5938	2.2307	2.3307	2.3686
	S-BIC	2.3686	2.5310	2.5921	2.2306	2.3304	2.3680
	P-AIC	2.3681	2.5302	2.5911	2.2306	2.3304	2.3680
	P-BIC	2.3685	2.5309	2.5926	2.2306	2.3306	2.3686
	Full model	2.3679	2.5297	2.5901	2.2305	2.3303	2.3677
Lower							
	S-AIC	1.6469	1.5059	1.4504	1.7774	1.6871	1.6533
	S-BIC	1.6466	1.5050	1.4488	1.7773	1.6869	1.6528
	P-AIC	1.6461	1.5042	1.4477	1.7772	1.6868	1.6527
	P-BIC	1.6465	1.5049	1.4492	1.7773	1.6870	1.6533
	Full model	1.6459	1.5037	1.4467	1.7772	1.6867	1.6524

5. Conclusions

This paper derives the asymptotic equivalence of the confidence intervals based on the model averaging estimator and the full model estimator under the framework of general parametric models and the semi-parametric VCPL models. More specifically, if the parameter of interest is a linear combination of regression coefficients, the intervals obtained from the FMA and full model are exactly equivalent. Hence, if the investigator's main interest is in interval estimation, the confidence interval based on the full model is recommended given its computational ease. Since our simulation findings suggest that FMA estimator generally has a smaller MSE than that obtained from the full model, alternative methods of developing confidence intervals based on the FMA estimator may exist which result in more efficient estimates. This is an issue that warrants further study. Moreover, whether this equivalence applies to other model frameworks provides another fruitful avenue for future research.

Appendix: Proofs

The proofs of Theorems 1 and 2 require the following technical conditions, which were also used in Fan and Huang (2005).

(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.

Proof of Theorem 1. By the Taylor series expansion,

$$\tilde{\mu}_{true} = \tilde{\mu}(\tilde{\theta}_0, \frac{\tilde{\delta}}{\sqrt{n}}) = \tilde{\mu}(\tilde{\theta}_0, 0) + \tilde{\mu}_{\tilde{\gamma}}^\top \frac{\tilde{\delta}}{\sqrt{n}} + o_P(\frac{1}{\sqrt{n}}),$$

and

$$\hat{\tilde{\mu}}_s = \tilde{\mu}(\hat{\tilde{\theta}}_s, \hat{\tilde{\gamma}}_s) = \tilde{\mu}(\tilde{\theta}_0, 0) + \begin{pmatrix} \tilde{\mu}_{\tilde{\theta}} \\ \pi_S \tilde{\mu}_{\tilde{\gamma}} \end{pmatrix}^\top \begin{pmatrix} \hat{\tilde{\theta}}_s - \tilde{\theta}_0 \\ \hat{\tilde{\gamma}}_s \end{pmatrix} + o_P(\frac{1}{\sqrt{n}})$$

Thus,

$$\begin{aligned}
&\hat{\tilde{\mu}}_s - \tilde{\mu}_{true} \\
&= \begin{pmatrix} \tilde{\mu}_{\tilde{\theta}} \\ \pi_S \tilde{\mu}_{\tilde{\gamma}} \end{pmatrix}^\top \begin{pmatrix} \hat{\tilde{\theta}}_s - \tilde{\theta}_0 \\ \hat{\tilde{\gamma}}_s \end{pmatrix} - \tilde{\mu}_{\tilde{\gamma}}^\top \frac{\delta}{\sqrt{n}} + o_P(\frac{1}{\sqrt{n}}) \\
&= \begin{pmatrix} \tilde{\mu}_{\tilde{\theta}} \\ \pi_S \tilde{\mu}_{\tilde{\gamma}} \end{pmatrix}^\top \begin{pmatrix} I_p & C_{ns} \\ 0_{r\times q} & (\pi_S A_n \pi_s^\top)^{-1} \pi_S A_n \end{pmatrix} \begin{pmatrix} \hat{\tilde{\theta}}_{\text{full}} - \tilde{\theta}_0 \\ \hat{\tilde{\gamma}}_{\text{full}} \end{pmatrix} - \tilde{\mu}_{\tilde{\gamma}}^\top \frac{\delta}{\sqrt{n}} + o_P(\frac{1}{\sqrt{n}}).
\end{aligned}$$

Expending the matrix product, we obtain that

$$\sqrt{n}(\hat{\tilde{\mu}}_s - \tilde{\mu}_{true}) = \tilde{\mu}_{\tilde{\theta}}^\top [M_{n1} + B_{11}^{-1} B_{12}(M_{n2} - \delta)] + \omega^\top \Big(\delta - A_n^{-1/2} H_s A_n^{1/2} M_{n2}\Big).$$

Using the similar approach in Fan and Huang (2005), we can show that

$$\begin{pmatrix} M_{n1} \\ M_{n2} \end{pmatrix} = \sqrt{n} \begin{pmatrix} \hat{\tilde{\theta}}_{\text{full}} - \tilde{\theta}_0 \\ \hat{\tilde{\gamma}}_{\text{full}} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \sim N\Bigg(\begin{pmatrix} 0 \\ \delta \end{pmatrix}, B^{-1} \Bigg).$$

Based on the Continuous mapping theorem and Slutsky Theorem, $\sqrt{n}(\hat{\tilde{\mu}}_s - \tilde{\mu}_{true})$ will converge in distribution to $\tilde{\Lambda}_s$.

□

Proof of Theorem 2. From definition of the FMA estimator in equation (17), we have

$$\sqrt{n}(\hat{\tilde{\mu}} - \tilde{\mu}_{\text{true}}) = \sum_S \tilde{c}(S|M_{n2}) \times \sqrt{n}(\hat{\tilde{\mu}}_S - \tilde{\mu}_{\text{true}}).$$

From the proof of Theorem 1, $\hat{\tilde{\mu}}_S - \tilde{\mu}_{\text{true}}$ is a linear function of M_{n1} and M_{n2}. Since $c(S|\cdot)$ is almost surely continuous, the equation above is an almost surely continuous function of M_{n1} and M_{n2}. Applying the Continuous mapping theorem, Slutsky theorem, and Theorem 1, we then obtain the required result. □

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory,* (B. N. Petrov and F. Cźaki, eds), pp. 267–281. Budapest: Akademiai Kaidó.

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

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.* New York: Cambridge University Press.

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

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

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.

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

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.

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

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. R., Powell, O., and Prufer, P. (2010). A comparison of two averaging techniques with an application to growth empirics. *Journal of Econometrics* 154:139–153.

Magnus, J. R., Wan, A. T. K., and Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the hong kong housing market. *Computational Statistics and Data Analysis* 55:1331–1341.

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

Schomaker, M., Wan, A. T. K., and Heumann, C. (2010). Frequentist model averaging with missing observations. *Computational Statistics and Data Analysis* 54:3336–3347.

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

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

Wan, A. T. K. and Zhang, X. (2009). On the use of model averaging in tourism research. *Annals of Tourism Research* 36:525–532.

Wan, A. T. K. and Zou, G. (2003). Optimal critical values of pre-tests when estimating the regression error variance: analytical findings under a general loss structure. *Journal of Econometrics* 114:165–196.

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

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* 96:574–586.

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

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. Z. (2012). Focused information criteria, model selection and model averaging in a tobit model with a non-zero threshold. *Journal of Business and Economic Statistics*, pages to appear.