




A Mallows-Type Model Averaging Estimator for the Varying-Coefficient Partially Linear Model

Rong Zhu, Alan T. K. Wan, Xinyu Zhang & Guohua Zou


To cite this article: Rong Zhu, Alan T. K. Wan, Xinyu Zhang & Guohua Zou (2019) A Mallows-Type Model Averaging Estimator for the Varying-Coefficient Partially Linear Model, Journal of the American Statistical Association, 114:526, 882-892, DOI: [10.1080/01621459.2018.1456936](https://doi.org/10.1080/01621459.2018.1456936)

To link to this article: <https://doi.org/10.1080/01621459.2018.1456936>

 View supplementary material [↗](#)

 Accepted author version posted online: 18 May 2018.
Published online: 06 Aug 2018.

 Submit your article to this journal [↗](#)

 Article views: 546

 View Crossmark data [↗](#)

 Citing articles: 1 View citing articles [↗](#)



A Mallows-Type Model Averaging Estimator for the Varying-Coefficient Partially Linear Model

Rong Zhu^{a,b}, Alan T. K. Wan^c, Xinyu Zhang^a, and Guohua Zou^d

^aAcademy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; ^bUniversity of Chinese Academy of Sciences, Beijing, China; ^cDepartment of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong; ^dSchool of Mathematical Sciences, Capital Normal University, Beijing, China

ABSTRACT

In the last decade, significant theoretical advances have been made in the area of frequentist model averaging (FMA); however, the majority of this work has emphasized parametric model setups. This article considers FMA for the semiparametric varying-coefficient partially linear model (VCPLM), which has gained prominence to become an extensively used modeling tool in recent years. Within this context, we develop a Mallows-type criterion for assigning model weights and prove its asymptotic optimality. A simulation study and a real data analysis demonstrate that the FMA estimator that arises from this criterion is vastly preferred to information criterion score-based model selection and averaging estimators. Our analysis is complicated by the fact that the VCPLM is subject to uncertainty arising not only from the choice of covariates, but also whether the covariate should enter the parametric or nonparametric parts of the model. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received January 2017

Revised March 2018

KEYWORDS

Asymptotic optimality; Heteroscedasticity; Mallows criterion; Model averaging; Varying-coefficient partially linear model

1. Introduction

The partially linear model (PLM) introduced by Engle et al. (1986) and Speckman (1988) is among the most popular semiparametric models in statistics. The PLM postulates that the response variable depends on some covariates in a parametric linear fashion and on other covariates nonparametrically. There have been widespread applications of the PLM in biomedicine, economics, and finance. See Härdle, Liang, and Gao (2000) for a review. Li et al. (2002) introduced the varying coefficient partially linear model (VCPLM) that replaces the general multi-dimensional nonparametric function in the PLM by a “varying coefficient function” (Cleveland, Grosse, and Shyu 1991; Hastie and Tibshirani 1993), which approximates the unknown nonparametric function by a conditional linear model, with the coefficients being low-dimensional smooth functions of certain covariates called the effect modifiers. Because only low-dimensional nonparametric functions are estimated, the VCPLM is less prone to the curse of dimensionality compared to the PLM. It is also more flexible in allowing interactions between a covariate and an unknown function through the effect modifiers.

An extensive body of literature focusing on the theory of the VCPLM has been developed. Li et al. (2002) proposed a local least-square method for estimating the smooth coefficient functions. They showed that the resultant estimators are consistent and asymptotically normal. Zhang, Lee, and Song (2002) developed a local polynomial regression-based estimation procedure, and Xia, Zhang, and Tong (2004) proposed an alternative semilocal estimation method that has the

advantage of reducing bias. Xia, Zhang, and Tong (2004) also developed a covariate selection procedure for the VCPLM. Ahmad, Leelahanon, and Li (2005) and Fan and Huang (2005) suggested two estimation approaches based on nonparametric series and profile least-square estimation, respectively. Both approaches yield semiparametric efficient estimators for the parametric components under the assumption of conditional homoscedasticity. Fan and Huang (2005) also considered inference, and showed that the Wilks phenomenon (Fan, Zhang, and Zhang 2001; Fan and Jiang 2007) holds for the profile likelihood ratio statistic, implying that its distribution is independent of unknown parameters. Wang, Zhu, and Zhou (2009) considered the VCPLM in the context of quantile regression. Zhao and Xue (2010) focused on covariate selection for the VCPLM when the covariates are measured with errors. They further proposed a bias-corrected covariate selection procedure by combining the basis function approximation with shrinkage estimation.

Qualifying the utility of covariates is an essential aspect in the application of statistical models, and the VCPLM is no exception. The traditional approach is to first estimate multiple competing models, each containing a different combination of covariates, then examine the fit of the models, and finally drawing inference from the best-fitting model, ignoring the alternative estimates and the uncertainty arising from the model selection process. Several studies, including Xia, Zhang, and Tong (2004) and Zhao and Xue (2010) mentioned above, have considered covariate selection for the VCPLM. Past research has shown that model selection underestimates the true variability of the model, and thereby results in overconfident

decision-making (Hjort and Claeskens 2003). In this article, we consider model averaging as an alternative to model selection within the VCPLM. Model averaging assigns weights to different models. These weights are then used to produce average estimates of the unknown parameters and functions. There is ample evidence indicating that model averaging frequently yields more accurate predictions of the target variable than model selection. In a well-cited article, Hansen (2007) developed a model weighting method for linear regressions based on a minimization of the Mallows criterion, and established an asymptotic optimality for the resultant model average estimator. Building on this work, Liu and Okui (2013) developed a heteroscedasticity-robust variant of Hansen’s (2007) Mallows model average estimator. Another common approach to model averaging is to assign weights based on information criterion scores obtained from different models, as in Buckland, Burnham, and Augustin (1997), Hjort and Claeskens (2006), and Zhang, Wan, and Zhou (2012).

The primary object of the current article is to develop a Mallows-type model averaging criterion for the VCPLM, to which the FMA literature has paid only scant attention to-date. Focusing on a VCPLM with measurement errors, Wang, Zou, and Wan (2012) derived the asymptotic distribution of model averaging estimators of the unknowns. The analysis of Wang, Zou, and Wan (2012) is based on the local misspecification framework introduced by Hjort and Claeskens (2003), which assumes that the true values of the auxiliary parameters in the model are in a \sqrt{n} -shrinking neighborhood of zero. Under this framework, for regular models or the family of models with local asymptotic normality, all maximum likelihood estimators are \sqrt{n} -consistent and have squared bias and variance of order $O(n^{-1})$. These properties facilitate the analysis of an estimator’s asymptotic behavior, but at the cost of assuming a model framework whose realism is subject to considerable criticism. Also, Wang, Zou, and Wan’s (2012) analysis considered only existing weighting methods based on information scores and did not propose any new method with superior properties.

Unlike Wang, Zou, and Wan (2012), our analysis is not based on the local misspecification framework; therefore, the validity of our results does not depend on the above-mentioned properties of the maximum likelihood estimators that arise from the local misspecification assumption. In addition, we allow for heteroscedasticity in the model’s errors. Our initial setup assumes a known covariance matrix of errors. Under this setup, we propose a weight choice criterion based on a minimization of an unbiased estimator of the expected predictive squared error (up to a constant) of the model average estimator. We prove that under certain regularity conditions, the weights resulting from this criterion lead to an asymptotically optimal model average estimator. We further show that when the unknown covariance matrix is estimated by a plugged-in estimator, the proposed criterion continues to yield an estimator that is asymptotically optimal. It is instructive to mention that our proof of optimality does not follow the methods of Hansen (2007) and Zhang and Wang (2018) who considered linear parametric and partial linear setups, respectively. Their proof techniques cannot be used if a varying coefficient nonparametric component is present. As well, there are two layers of uncertainty for the VCPLM: the first concerns the choice of covariates, while the second concerns whether the covariate should enter the parametric or nonparametric parts of the model. These features

of the VCPLM significantly complicate the analysis and pose formidable technical challenges with regard to establishing an asymptotic theory for the FMA estimator.

The remainder of this article is organized as follows. Section 2 contains a description of the VCPLM and methods for estimation. Section 3 introduces the Mallows-type weight choice criterion and the varying coefficient partially linear model average (VCPLMA) estimator that arises from this criterion. The asymptotic optimality of the VCPLMA estimator is established in Section 4. Section 5 compares the finite sample properties of the VCPLMA estimator with several information criterion-based model selection and averaging estimators. A real data example is considered in Section 6. Section 7 contains some concluding remarks. An Appendix contains the sketches of the proofs of theorems. Detailed proofs are given in an online supplemental file.

2. Model Set-Up and Parametric Estimation

The VCPLM is described by the following equation:

$$y_i = \mu_i + \epsilon_i = x_i^T \beta + z_i^T \alpha(t_i) + \epsilon_i = \sum_{k=1}^K x_{ik} \beta_k + \sum_{r=1}^R z_{ir} \alpha_r(t_i) + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (1)$$

where y_i is the response variable, $x_i = (x_{i1}, x_{i2}, \dots, x_{iK})^T$ and $z_i = (z_{i1}, z_{i2}, \dots, z_{iR})^T$ are covariates, $\beta = (\beta_1, \beta_2, \dots, \beta_K)^T$ is an unknown coefficient vector associated with x_i , $\alpha(\cdot) = (\alpha_1(\cdot), \alpha_2(\cdot), \dots, \alpha_R(\cdot))^T$ is an unknown coefficient function vector associated with z_i , t_i is the “effect modifier,” a univariate covariate different from x_i and z_i , and ϵ_i is a disturbance term with conditional mean $E(\epsilon_i | x_i, z_i, t_i) = 0$ and variance $E(\epsilon_i^2 | x_i, z_i, t_i) = \sigma_i^2$. Model (1) permits interactions between t_i and z_i in such a way that the effects of z_i vary over different levels of t_i ; a different level of t_i is also associated with a different linear model. In the model averaging literature, the true model is commonly assumed to be infinite dimensional (Hansen 2007; Lu and Su 2015). We make the same assumption here, and let K and R go to infinity.

Equation (1) may be expressed equivalently in matrix form as

$$Y = \mu + \epsilon = X\beta + S + \epsilon, \quad (2)$$

where $Y = (y_1, y_2, \dots, y_n)^T$ is an $n \times 1$ vector of the dependent variable, $X = (x_1, x_2, \dots, x_n)^T$ is an $n \times K$ covariate matrix, $S = (z_1^T \alpha(t_1), z_2^T \alpha(t_2), \dots, z_n^T \alpha(t_n))^T$, $Z = (z_1, z_2, \dots, z_n)^T$ is an $n \times R$ covariate matrix, $\Psi = (t_1, t_2, \dots, t_n)^T$ is an $n \times 1$ vector of the effect modifier, $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T = X\beta + S$ is an $n \times 1$ function vector of X, Z , and Ψ , and $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ is an $n \times 1$ vector of random disturbances with $E(\epsilon | X, Z, \Psi) = 0$ and $E(\epsilon \epsilon^T | X, Z, \Psi) = \Omega = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$.

We use M candidate models to approximate the true data-generating process of Y , where M is allowed to diverge to infinity as $n \rightarrow \infty$. The m th candidate (or approximating) VCPLM includes k_m of K regressors in x_i and r_m of R regressors in z_i . Thus, the m th approximating model is

$$Y = \mu_{(m)} + \epsilon_{(m)} = X_{(m)} \beta_{(m)} + S_{(m)} + \epsilon_{(m)}, \quad m = 1, 2, \dots, M, \quad (3)$$

where $X_{(m)}$ is an $n \times k_m$ matrix containing k_m columns of X with full column rank, $\epsilon_{(m)}$ is the error term of the m th approximating model, $S_{(m)} = (z_{(m)1}^\top \alpha_{(m)}(t_1), z_{(m)2}^\top \alpha_{(m)}(t_2), \dots, z_{(m)n}^\top \alpha_{(m)}(t_n))^\top$, $Z_{(m)} = (z_{(m)1}, z_{(m)2}, \dots, z_{(m)n})^\top$ is an $n \times r_m$ matrix containing r_m columns of Z with full column rank (i.e., the m th candidate model contains k_m covariates in X and r_m covariates in Z), $\beta_{(m)} = (\beta_{(m)1}, \beta_{(m)2}, \dots, \beta_{(m)k_m})^\top$ is a $k_m \times 1$ unknown coefficient vector, and $\alpha_{(m)}(t) = (\alpha_{(m)1}(t), \alpha_{(m)2}(t), \dots, \alpha_{(m)r_m}(t))^\top$ is an $r_m \times 1$ unknown coefficient function vector that varies with t . Following the convention in the model averaging literature, we assume that all candidate models are incorrect, and at best, approximations to the true data-generating process.

Remark 1. Motivated by the common aphorism in statistics that all models are wrong but some are useful (Box 1976; Box and Draper 1987), studies of model averaging commonly assume that no candidate model is the true model. See Hansen (2007), Wan, Zhang, and Zou (2010), Hansen and Racine (2012), and Liu and Okui (2013), where the assumption is made explicitly. It is an important assumption to ensure the validity of our as well as existing results on the asymptotic optimality of model average estimators as the conditions on ξ_n stated under Conditions (C.2) and (C.9) in Section 4 cannot hold if the true model is among the candidate set.

We estimate $\beta_{(m)}$ and $\alpha_{(m)}(t)$ by the profile least-square method (Fan and Huang 2005; You and Chen 2006) described as follows. For any t in the neighborhood of t_0 , let $\alpha_{(m)j}(t)$ be approximated by the following linear function:

$$\begin{aligned} \alpha_{(m)j}(t) &\approx \alpha_{(m)j}(t_0) + \alpha'_{(m)j}(t_0)(t - t_0) \\ &\equiv a_{(m)j} + b_{(m)j}(t - t_0), \quad j = 1, 2, \dots, r_m. \end{aligned}$$

If $\beta_{(m)}$ is known, then $a_{(m)j}$ and $b_{(m)j}$ are solutions to the following weighted local least-square criterion:

$$\begin{aligned} \min_{a_{(m)}, b_{(m)}} \sum_{i=1}^n [y_i - x_{(m)i}^\top \beta_{(m)} - z_{(m)i}^\top \{a_{(m)} + b_{(m)}(t_i - t_0)\}]^2 \\ \times K_{h_m}(t_i - t_0), \end{aligned} \tag{4}$$

where $a_{(m)} = (a_{(m)1}, a_{(m)2}, \dots, a_{(m)r_m})^\top$, $b_{(m)} = (b_{(m)1}, b_{(m)2}, \dots, b_{(m)r_m})^\top$, $K_{h_m}(\cdot) = K(\cdot/h_m)/h_m$, $K(\cdot)$ is a kernel function, and h_m is a bandwidth. The solutions may be expressed as

$$\begin{aligned} (\hat{a}_{(m)1}(t), \dots, \hat{a}_{(m)r_m}(t), h_m \hat{b}_{(m)1}(t), \dots, h_m \hat{b}_{(m)r_m}(t))^\top \\ = \{D_{(m)t}^\top W_{(m)t} D_{(m)t}\}^{-1} D_{(m)t}^\top W_{(m)t} (Y - X_{(m)} \beta_{(m)}), \end{aligned} \tag{5}$$

where $W_{(m)t} = \text{diag}\{K_{h_m}(t_1 - t), K_{h_m}(t_2 - t), \dots, K_{h_m}(t_n - t)\}$, $D_{(m)t} = \begin{pmatrix} z_{(m)1}^\top & \frac{t_1-t}{h_m} z_{(m)1}^\top \\ \vdots & \vdots \\ z_{(m)n}^\top & \frac{t_n-t}{h_m} z_{(m)n}^\top \end{pmatrix}_{n \times 2r_m} = (Z_{(m)}; \Lambda_{(m)t} Z_{(m)}) =$

$(Z \Pi_{2m}^\top; \Lambda_{(m)t} Z \Pi_{2m}^\top)$, $\Lambda_{(m)t} = \text{diag}\{\frac{t_1-t}{h_m}, \frac{t_2-t}{h_m}, \dots, \frac{t_n-t}{h_m}\}$, and Π_{2m} is an $r_m \times R$ selection matrix.

Remark 2. In theory, one can use different h_m 's for different candidate models, $m = 1, 2, \dots, M$. However, this requires tremendous computational efforts especially when M is large.

In our numerical analysis, we use the relatively simple rule-of-thumb bandwidth selection method that results in a common bandwidth for all models. See Sections 5 and 6 for details.

Suppose that the models of interest are indexed by the subsets $\{U_1\}$ of $\{1, 2, \dots, K\}$. Let Π_{1m} be the selection matrix mapping $x_i = (x_{i1}, x_{i2}, \dots, x_{iK})^\top$ onto the subvector $x_{(m)i} = \Pi_{1m} x_i$ of components x_{ij} with $j \in U_1$. Hence, Π_{1m} is of size $k_m \times K$ with k_m being the size of U_1 . Similarly, let the subsets $\{U_2\}$ of $\{1, 2, \dots, R\}$ index the models of interest, and Π_{2m} be the projection matrix mapping $z_i = (z_{i1}, z_{i2}, \dots, z_{iR})^\top$ onto the subvector $z_{(m)i} = \Pi_{2m} z_i$ of components z_{ij} with $j \in U_2$. Hence, Π_{2m} is of size $r_m \times R$ with r_m being the size of U_2 .

Substituting $(\hat{a}_{(m)1}(t), \hat{a}_{(m)2}(t), \dots, \hat{a}_{(m)r_m}(t))$ in model (3) yields

$$y_i - \hat{y}_{(m)i} = (x_{(m)i} - \hat{x}_{(m)i})^\top \beta_{(m)} + \tilde{\epsilon}_i, \tag{6}$$

where $\tilde{\epsilon}_i$ is the random error different from ϵ_i ,

$$\hat{y}_{(m)i} = (z_{(m)i}^\top, 0^\top) \{D_{(m)t_i}^\top W_{(m)t_i} D_{(m)t_i}\}^{-1} D_{(m)t_i}^\top W_{(m)t_i} Y,$$

and

$$\hat{x}_{(m)i} = [(z_{(m)i}^\top, 0^\top) \{D_{(m)t_i}^\top W_{(m)t_i} D_{(m)t_i}\}^{-1} D_{(m)t_i}^\top W_{(m)t_i} X_{(m)}]^\top.$$

Denote

$$A_{(m)} = \begin{pmatrix} (z_{(m)1}^\top, 0^\top) \{D_{(m)t_1}^\top W_{(m)t_1} D_{(m)t_1}\}^{-1} D_{(m)t_1}^\top W_{(m)t_1} \\ \vdots \\ (z_{(m)n}^\top, 0^\top) \{D_{(m)t_n}^\top W_{(m)t_n} D_{(m)t_n}\}^{-1} D_{(m)t_n}^\top W_{(m)t_n} \end{pmatrix}_{n \times n}, \tag{7}$$

$\hat{X}_{(m)} = A_{(m)} X_{(m)}$, $\hat{Y}_{(m)} = A_{(m)} Y$, and $P_{(m)} = \hat{P}_{(m)} (I_n - A_{(m)}) + A_{(m)}$, where $\hat{P}_{(m)} = (I_n - A_{(m)}) X_{(m)} \{X_{(m)}^\top (I_n - A_{(m)})^\top (I_n - A_{(m)}) X_{(m)}\}^{-1} X_{(m)}^\top (I_n - A_{(m)})^\top$ is an $n \times n$ projection matrix. Fan and Huang (2005) showed that the least-square estimator of $\beta_{(m)}$ in (6) is

$$\begin{aligned} \hat{\beta}_{(m)} &= \{(X_{(m)} - \hat{X}_{(m)})^\top (X_{(m)} - \hat{X}_{(m)})\}^{-1} \\ &\quad \times (X_{(m)} - \hat{X}_{(m)})^\top (Y - \hat{Y}_{(m)}) \\ &= \{X_{(m)}^\top (I_n - A_{(m)})^\top (I_n - A_{(m)}) X_{(m)}\}^{-1} X_{(m)}^\top \\ &\quad \times (I_n - A_{(m)})^\top (I_n - A_{(m)}) Y. \end{aligned} \tag{8}$$

Under the m th candidate model, the estimator of μ is given by

$$\begin{aligned} \hat{\mu}_{(m)} &\equiv X_{(m)} \hat{\beta}_{(m)} + \hat{S}_{(m)} \\ &= X_{(m)} \hat{\beta}_{(m)} + A_{(m)} (Y - X_{(m)} \hat{\beta}_{(m)}) \\ &= (I_n - A_{(m)}) X_{(m)} \hat{\beta}_{(m)} + A_{(m)} Y \\ &= [(I_n - A_{(m)}) X_{(m)} \{X_{(m)}^\top (I_n - A_{(m)})^\top (I_n - A_{(m)}) X_{(m)}\}^{-1} X_{(m)}^\top \\ &\quad \times (I_n - A_{(m)})^\top (I_n - A_{(m)}) + A_{(m)}] Y \\ &= \{\hat{P}_{(m)} (I_n - A_{(m)}) + A_{(m)}\} Y \\ &= P_{(m)} Y. \end{aligned} \tag{9}$$

Equation (9) shows that $\hat{\mu}_{(m)}$ is linearly dependent on Y .

3. Model Averaging and Weight Choice Criterion

Let $w = (w_1, w_2, \dots, w_M)^T$ be a weight vector in the unit simplex of \mathbb{R}^M :

$$\mathcal{H}_n = \left\{ w \in [0, 1]^M : \sum_{m=1}^M w_m = 1 \right\}.$$

The model average estimator of μ is

$$\hat{\mu}(w) = \sum_{m=1}^M w_m \hat{\mu}_{(m)} = \sum_{m=1}^M w_m P_{(m)} Y = P(w) Y,$$

where $P(w) = \sum_{m=1}^M w_m P_{(m)}$. Define the squared error loss function and the corresponding conditional risk function as $L_n(w) = \|\hat{\mu}(w) - \mu\|^2$ and

$$\begin{aligned} R_n(w) &= E(L_n(w) | X, Z, \Psi) \\ &= E(\|(P(w) - I_n)\mu + P(w)\epsilon\|^2 | X, Z, \Psi) \\ &= \|(P(w) - I_n)\mu\|^2 + \text{tr}(P^T(w)P(w)\Omega), \end{aligned} \tag{10}$$

respectively. Our choice of w is based on the criterion

$$C_n(w) = \|Y - \hat{\mu}(w)\|^2 + 2\text{tr}(P(w)\Omega). \tag{11}$$

It can be readily shown that

$$\begin{aligned} E(C_n(w) | X, Z, \Psi) &= E(\|\mu + \epsilon - \hat{\mu}(w)\|^2 \\ &\quad + 2\text{tr}(P(w)\Omega) | X, Z, \Psi) \\ &= E(\|\mu - \hat{\mu}(w)\|^2 + \|\epsilon\|^2 \\ &\quad + 2\epsilon^T(\mu - \hat{\mu}(w)) \\ &\quad + 2\text{tr}(P(w)\Omega) | X, Z, \Psi) \\ &= E(L_n(w) | X, Z, \Psi) + \text{tr}(\Omega) \\ &\quad - 2E\{\text{tr}(P(w)\Omega) | X, Z, \Psi\} \\ &\quad + 2E\{\text{tr}(P(w)\Omega) | X, Z, \Psi\} \\ &= E(L_n(w) | X, Z, \Psi) + \text{tr}(\Omega). \end{aligned}$$

In other words, $C_n(w)$ is an unbiased estimator of the expected in-sample squared error loss plus a constant, which is similar to the Mallows criterion proposed by Hansen (2007). Thus, our criterion is a Mallows-type criterion.

The optimal weight vector is obtained by minimizing $C_n(w)$ over the weight set \mathcal{H}_n , that is,

$$\hat{w} = \arg \min_{w \in \mathcal{H}_n} C_n(w). \tag{12}$$

Given that (12) is a quadratic programming problem, the computation of the optimal weight vector is straightforward. We refer to the resultant estimator $\hat{\mu}(\hat{w})$ as the VCPLMA estimator.

In practice, the covariance matrix Ω is unknown and needs to be estimated. Following Hansen (2007), we estimate Ω based on the largest approximating model indexed by $M^* = \arg \max_{1 \leq m \leq M} \{k_m + r_m\}$, leading to the estimator

$$\hat{\Omega} = \text{diag}(\hat{\epsilon}_{(M^*)1}^2, \hat{\epsilon}_{(M^*)2}^2, \dots, \hat{\epsilon}_{(M^*)n}^2), \tag{13}$$

where $\hat{\epsilon}_{(M^*)i} = y_i - \hat{\mu}_{(M^*)i}$ and $\hat{\mu}_{(M^*)i}$ is the i th component of $\hat{\mu}_{(M^*)}$.

When Ω is replaced by $\hat{\Omega}$, the criterion in (11) changes accordingly to

$$\hat{C}_n(w) = \|Y - \hat{\mu}(w)\|^2 + 2\text{tr}(P(w)\hat{\Omega}), \tag{14}$$

which may be treated as a feasible counterpart of $C_n(w)$. Minimizing $\hat{C}_n(w)$ with respect to w leads to

$$\tilde{w} = \arg \min_{w \in \mathcal{H}_n} \hat{C}_n(w). \tag{15}$$

Substituting \tilde{w} in $\hat{\mu}(w)$ yields the VCPLMA estimator under the unknown Ω case.

4. Asymptotic Optimality of the VCPLMA Estimator

Let $\xi_n = \inf_{w \in \mathcal{H}_n} R_n(w)$, $\tilde{r} = \max_{1 \leq m \leq M} r_m$, $\tilde{h} = \max_{1 \leq m \leq M} h_m$, $h = \min_{1 \leq m \leq M} h_m$, $\bar{\lambda}(\cdot)$ and $\underline{\lambda}(\cdot)$ be the maximum and minimum singular values of a given matrix, respectively, and w_m^0 be an $M \times 1$ vector where the m th element is one and all other elements are zeros. The following regularity conditions are required for the VCPLMA estimator to achieve asymptotic optimality. All limiting processes, unless stated otherwise, correspond to $n \rightarrow \infty$.

Condition (C.1). For some fixed integer G ($1 \leq G < \infty$) and constant $\kappa < \infty$, $E(\epsilon_i^{4G} | x_i, z_i, t_i) \leq \kappa < \infty$, for all $i = 1, 2, \dots, n$, a.s.

Condition (C.2). $M \xi_n^{-2G} \tilde{r}^G \sum_{m=1}^M \{R_n(w_m^0)\}^G = o_p(1)$.

Condition (C.3). The random variable t_i has bounded support Δ . The density $f(\cdot)$ of t_i is continuous and bounded away from 0 on its support, which is at least twice differentiable. The function $\alpha_j(\cdot)$ is twice continuously differentiable in Δ for all $j = 1, 2, \dots, R$.

Condition (C.4). $\max_{1 \leq m \leq M} \max_{1 \leq i \leq n} z_{(m)i}^T z_{(m)i} = O_p(\tilde{r})$, $C_z \equiv E(z_j z_j^T)$ is nonsingular for all j ; as well, there exist constants κ_z , \underline{c}_z , and \bar{c}_z , such that $E\{(z_{li} z_{lk})^2\} \leq \kappa_z < \infty$, for all $l = 1, 2, \dots, n$; $i, k = 1, 2, \dots, R$, and $0 < \underline{c}_z \leq \min_{1 \leq m \leq M} \underline{\lambda}(\Pi_{2m} C_z \Pi_{2m}^T) \leq \max_{1 \leq m \leq M} \bar{\lambda}(\Pi_{2m} C_z \Pi_{2m}^T) \leq \bar{c}_z < \infty$.

Condition (C.5). The kernel function $K(\cdot)$ is a bounded symmetrical density with symmetrical and compact support $\text{supp}(K)$, $\tilde{h} = O(n^{-1/5})$ and $h = O(n^{-1/5})$.

Remark 3. Condition (C.1) places a restriction on the conditional moment of the error term. Condition (C.2) is analogous to Condition (8) of Wan, Zhang, and Zou (2010), which is widely used in studies of model averaging. See, for example, Liu and Okui (2013) and Ando and Li (2014). Condition (C.3) is an assumption related to the densities of t_i and its functions. It is similar to Conditions (C1) and (C4) of Wang, Zou, and Wan (2012). Condition (C.4) places a mild condition on Z ; it is related to the norm of $z_{(m)i}$, the moment of $z_{li} z_{lk}$, and the singular value of the expectation of $z_j z_j^T$. Condition (C.5) is a common assumption of the kernel function and bandwidth.

The next theorem gives the asymptotic optimality of the VCPLMA estimator when Ω is known.

Theorem 1. Let Conditions (C.1)–(C.5) hold. Then

$$\frac{L_n(\hat{w})}{\inf_{w \in \mathcal{H}_n} L_n(w)} \xrightarrow{P} 1. \tag{16}$$

Theorem 1 shows that the VCPLMA estimator based on \hat{w} is asymptotically optimal in that it leads to a squared error loss that is asymptotically identical to that of the infeasible best possible model average estimator. The Appendix provides a sketch of the proof of **Theorem 1**. The detailed proof is available in the online supplemental file.

Let $\rho_{ii}^{(m)}$ be the i th diagonal element of $P_{(m)}$ and $\tilde{k} = \max_{1 \leq m \leq M} k_m$. When Ω is estimated by $\hat{\Omega}$ given in (13), provided that the following additional conditions are satisfied, it can be shown that the VCPLMA estimator based on \tilde{w} shares the same asymptotic optimality as the corresponding estimator based on \hat{w} described in **Theorem 1**.

Condition (C.6). There exists a constant c such that $|\rho_{ii}^{(m)}| \leq cn^{-1}|tr(P_{(m)})|$ for all $m \in \{1, 2, \dots, M\}$.

Condition (C.7). $\mu^T \mu / n = O(1)$, a.s..

Condition (C.8). $n^{-1}\tilde{r}^2\tilde{k}^2 = O(1)$ and $n^{-1}h^{-2}\tilde{r}^3 = O(1)$.

Condition (C.9). $\xi_n^{-1}\tilde{r}^3/2\tilde{k} = o_p(1)$ and $\xi_n^{-1}h^{-1}\tilde{r}^2 = o_p(1)$.

Remark 4. Condition (C.6) is commonly used to ensure the asymptotic optimality of cross-validation. See, for example, Andrews (1991) and Hansen and Racine (2012). Condition (C.7) is about the sum of the elements of μ and is commonly used in the context of linear regression. See, for example, Wan, Zhang, and Zou (2010) and Liang et al. (2011). Condition (C.8) has two parts—the first part places a restriction on the rate of increase of $\tilde{r}^2\tilde{k}^2$ as $n \rightarrow \infty$, while the second part is about the relationship between the bandwidth and \tilde{r}^3 . Condition (C.9) implies that ξ_n increases at a rate faster than $\tilde{r}^3/2\tilde{k}$ and $h^{-1}\tilde{r}^2$.

Theorem 2. Suppose that Conditions (C.1)–(C.9) hold. Then as $n \rightarrow \infty$, we have

$$\frac{L_n(\tilde{w})}{\inf_{w \in \mathcal{H}_n} L_n(w)} \xrightarrow{P} 1. \tag{17}$$

Theorem 2 shows that **Theorem 1** remains valid when Ω is replaced by $\hat{\Omega}$. The Appendix provides a sketch of the proof of **Theorem 2**. A detailed proof is available in the online supplemental file.

Remark 5. Our criterion in (11) contains the term $tr(P(w)\Omega)$. Similar to the approach of Liu and Okui (2013), instead of treating Ω in isolation, we treat $tr(P(w)\Omega)$ as one entity and estimate it by $\sum_{i=1}^n \hat{e}_i^2 p_{ii}(w)$, where $\hat{e}_i = y_i - \hat{\mu}_{(M^*)i}$ and $p_{ii}(w)$ is the i th diagonal element of $P(w)$.

5. A Simulation Study

This section is devoted to a comparison of the finite-sample performance of the VCPLMA estimator that arises from the proposed Mallows-type weight choice method with several existing information criterion-based model selection and

averaging methods in a Monte Carlo study. We will begin with a description of the experimental design, followed by the estimation procedure and results.

5.1. Experimental Design

We assume that the true data-generating process of y_i is given by

$$y_i = \mu_i + \epsilon_i = \sum_{k=1}^{200} x_{ik}\beta_k + \sum_{r=1}^{200} z_{ir}\alpha_r(t_i) + \epsilon_i,$$

where the observations of $x_i = (x_{i1}, x_{i2}, \dots, x_{i200})^T$ and $z_i = (z_{i1}, z_{i2}, \dots, z_{i200})^T$ are generated from a multivariate normal distribution with mean 0 and covariance Σ , with the ij th element of Σ being Σ_{ij} , $t_i \stackrel{iid}{\sim} U(0, 1)$, and $\epsilon_i \sim N(0, \eta^2(x_{i2}^2 + 0.01))$. We vary η such that $R^2 = \text{var}(\mu_i)/\text{var}(y_i)$ varies between 0.1 and 0.9, where $\text{var}(\mu_i)$ and $\text{var}(y_i)$ are the variances of μ_i and y_i , respectively. We set $n = 50, 100, 200$, and 400, and consider six experimental designs that differ in terms of $\beta_k, \alpha_r(t_i), M$ and the covariances between the elements of x_i and z_i . They are showed in **Table 1**. In the table, Σ_{ij} denotes the ij th element of Σ , the covariance matrices of x_i and z_i , and the function $\text{INT}(\cdot)$ returns the value of the figure inside the bracket rounded to the nearest integer. Thus, for Designs 1 and 2, $M = 11, 14, 18$, and 22 for $n = 50, 100, 200$, and 400, respectively. The value of M for Designs 3–6 arises from the condition that at least one of the variables in $\{x_{i1}, x_{i2}, x_{i3}, z_{i1}\}$ must be a covariate in either one of the two component but not in both. Hence, there are $M = \binom{4}{1}(2^3 - 1) + \binom{4}{2}(2^2 - 1) + \binom{4}{3} = 50$ candidate models.

5.2. Estimation and Comparison

We use the Epanechnikov kernel $K(v) = \frac{3}{4}(1 - v^2)I_{\{|v| \leq 1\}}$ and set the bandwidth h_m to $\text{std}(t)n^{-1/5}$, which is the optimal bandwidth choice based on the rule-of-thumb method, for all $m = 1, 2, \dots, M$, where $\text{std}(t)$ is the sample standard deviation of $\{t_i\}_{i=1}^n$ (see **Remark 2**). For Designs 1 and 2, where covariate uncertainty exists only in the parametric component, Ω is estimated based on the M th candidate model containing the largest number of covariates. Under Designs 3–6, it is estimated based on the model with the largest number of covariates across the parametric and nonparametric parts.

We compare the finite sample performance of the VCPLMA estimator with the AIC- and BIC-based model selection and averaging estimators. The AIC and BIC scores for the m th candidate model are $\text{AIC}_m = \log(\hat{\sigma}_m^2) + 2n^{-1}\text{tr}(P_{(m)})$ and $\text{BIC}_m = \log(\hat{\sigma}_m^2) + n^{-1}\text{tr}(P_{(m)}) \log(n)$, respectively, where $\hat{\sigma}_m^2 = n^{-1}\|y - \hat{\mu}_{(m)}\|^2$. These two criteria each select a model that corresponds to the smallest of their respective scores. Buckland, Burnham, and Augustin (1997) suggested a weight choice for model averaging based on the following smoothed-version of the AIC and BIC: $\text{SAIC}_m = \exp(-\text{AIC}_m/2) / \sum_{l=1}^M \exp(-\text{AIC}_l/2)$ and $\text{SBIC}_m = \exp(-\text{BIC}_m/2) / \sum_{l=1}^M \exp(-\text{BIC}_l/2)$. Due to its ease of use, the SAIC and SBIC weight choice methods have been

Table 1. Summary of experimental designs for simulation study.

Design	β_k	$\alpha_r(t_i)$	Σ_{ij}	M	Covariate set
1	$\frac{1}{k^2}$	$\frac{\sin(2\pi r t_i)}{r}$	$(1/2)^{ i-j }$	$\text{INT}(3n^{\frac{1}{3}})$	Assume that z_{j1} is the only covariate included in the nonparametric part resulting in a common nonparametric structure of $z_{j1}\alpha_1(t_i)$ for all candidate models. Covariates in the parametric part are drawn from the set $\{x_{j1}, x_{j2}, \dots, x_{jm}\}$, with the m th candidate model containing the first m variables in the set, that is, the candidate models are nested.
2	$\frac{1}{k^2}$	$\frac{\sin(2\pi r t_i)}{r}$	$(1/2)^{ i-j }$	$\text{INT}(3n^{\frac{1}{3}})$	Identical to Design 1 except that all models contain the same two covariates z_{j1} and z_{j2} in the nonparametric part, resulting in a common nonparametric structure of $z_{j1}\alpha_1(t_i) + z_{j2}\alpha_2(t_i)$ for all models.
3	$\frac{1}{k^2}$	$\frac{\sin(2\pi r t_i)}{r}$	$(1/2)^{ i-j }$	50	The covariate set contains $\{x_{j1}, x_{j2}, x_{j3}, z_{j1}\}$. A candidate model must contain at least one covariate from the above set in either the parametric or nonparametric components but not in both.
4	$\frac{1}{k^{3/2}}$	$\frac{\sin(2\pi r t_i)}{r}$	$(1/2)^{ i-j }$	50	Identical to Design 3
5	$\frac{1}{k^2}$	$t_i e^{-r t_i}$	$(1/2)^{ i-j }$	50	Identical to Design 3
6	$\frac{1}{k^2}$	$\frac{\sin(2\pi r t_i)}{r}$	1 if $i = j$, 0 if $i \neq j$	50	Identical to Design 3

used extensively in the FMA literature. Examples are Hjort and Claeskens (2006) and Zhang, Wan, and Zhou (2012).

Our evaluation of the performance of estimators is based on the following sample mean squared error (MSE) of the response variable:

$$\text{MSE}^{(d)} = \frac{1}{nD} \sum_{d=1}^D \|\hat{\mu}^{(d)} - \mu^{(d)}\|^2, \tag{18}$$

where $D = 500$ is the number of replications and d indexes the d th simulation trial.

5.3. Results

As the results produced are similar between Designs 1 and 2, and among Designs 3–6, we only present the results corresponding to Designs 1, 3, and 6, which are given in Figures 1–3, respectively. One remarkable aspect of the results is that over a very large region of the parameter space, the VCPLMA estimator is found to deliver vastly more accurate outcomes than the other four competing estimators, including the FMA estimators based on the SAIC and SBIC averaging methods. The superiority of the VCPLMA estimator over the others is more pronounced under Designs 3–6, where covariate uncertainty exists in both the parametric and nonparametric parts, than under Designs 1 and 2, where the inclusion of covariates is uncertain only for the parametric part. The MSE reduction relative to model selection achieved by the VCPLMA estimator is also more substantial when R^2 is small than when it is large. Typically, a small R^2 is associated with a high noise content in the model’s disturbances. Under this situation, it is often difficult to identify a single best model, leading to model selection estimators exhibiting very unstable and often poor results. Averaging, on the other hand, smoothes across all candidate models, and thus shields against choosing and subsequently relying on a very bad model. The exact opposite explains why selection can sometimes outperform averaging when R^2 is large. However, in this case often one single model can take up an overwhelming proportion of model weights in the model

average, especially when n is large, yielding a model average estimator with an MSE that is effectively indistinguishable from that obtained based on the model selection approach. Except when R^2 is large, the VCPLMA estimator is found to produce more accurate outcomes than the SAIC and SBIC estimators, which in turn outperform their respective model selection counterparts. We are especially encouraged by the marked superiority observed for the VCPLMA estimator under Designs 3–6, which suggests that the VCPLMA estimator is most useful when covariate uncertainty exists in both the parametric and nonparametric parts of the model, a phenomenon commonly encountered in practice. Other things being equal, a large n , which reduces the noise level of the model, generally plays in model selection’s favor; as n increases, all estimators enjoy a reduction in MSE. Recall that the optimality of the VCPLMA estimator does not depend on the ability to include the true model in the candidate set. We think that the good showing of the estimator in finite samples may be attributed to this merit.

6. Empirical Application

We now apply the proposed VCPLMA methodology to a real dataset that contains observations on aged patients in 36 nursing homes in San Diego, CA, collected between 1980 and 1982. The same data were used by Fan, Lin, and Zhou (2006), Morris, Norton, and Zhou (1994), and Xie, Wan, and Zhou (2015) in their studies. The dependent variable of interest, y , is the natural logarithm of the number of days the patient stayed in a nursing home. The covariates include x_1 , an indicator equal to 1 if the patient received medical treatment at the nursing home and 0 otherwise; x_2 , which equals 1(0) if the patient is male(female); x_3 , which equals 1(0) if the patient is married(not married); x_4 , which represents health status, with larger x_4 indicating worse health conditions; $t = (\text{age} - 64)/(102 - 64)$ is the normalized age of the patients in the sample, and age lies between 65 and 102. We treat t as the effect modifier. The original sample contains 1601 observations including 332 censored observations. Our analysis is based on the remaining 1269 uncensored observations.

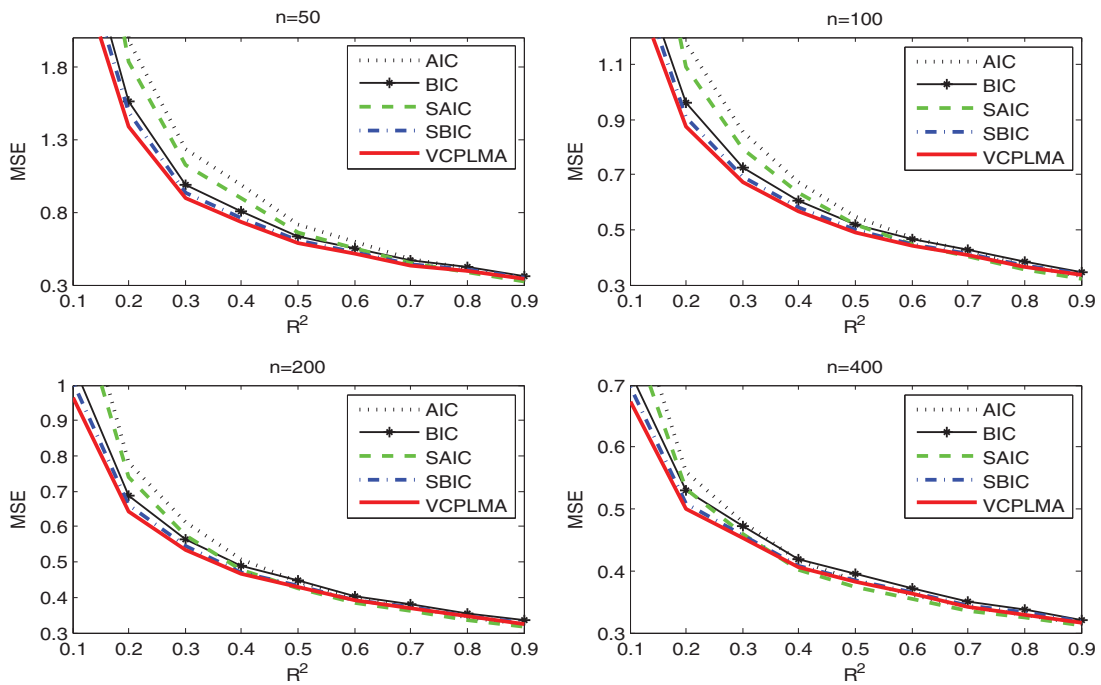


Figure 1. MSE comparisons under Design 1.

We have no prior knowledge of which of $x_1, x_2, x_3,$ and x_4 to include in the model. All of our candidate models contain no fewer than one covariate in each of the parametric and nonparametric parts, with no overlapping covariates in the two. With four covariates, this results in $M = 50$ candidate models. We randomly divide the data into a training sub-sample and a test subsample. Let n_0 be the number of observations in the training subsample. We set n_0 to 600, 700, 800, 900, 1000, 1100, and 1200. Based on the estimated model, we forecast the

remaining $n_1 = n - n_0$ observations of y in the corresponding test subsample. Our performance metric of predictive efficiency is the normalized mean squared prediction error (NMSPE), obtained by dividing the MSPE of a given estimator by the MSPE of the infeasible optimal estimator, which is the estimator based on one of the $M = 50$ models that yields the smallest MSPE across the n_1 test observations. We repeat this process $D = 500$ times, and calculate the mean and the median of the normalized NMSPEs of the five methods across the replications.

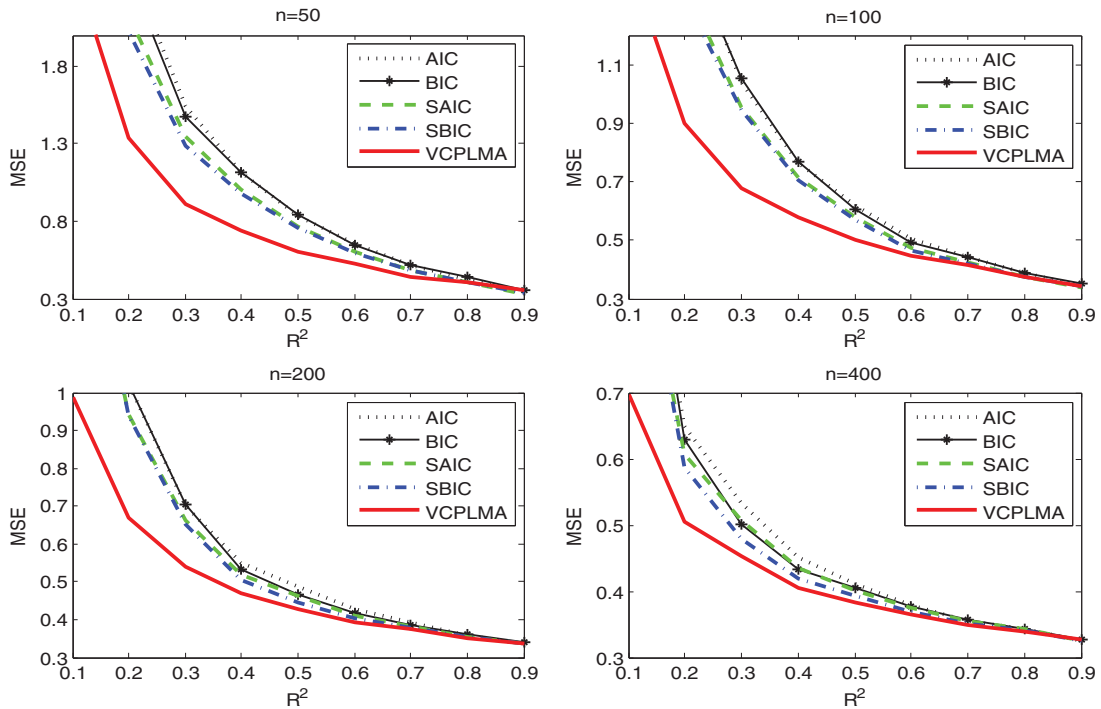


Figure 2. MSE comparisons under Design 3.

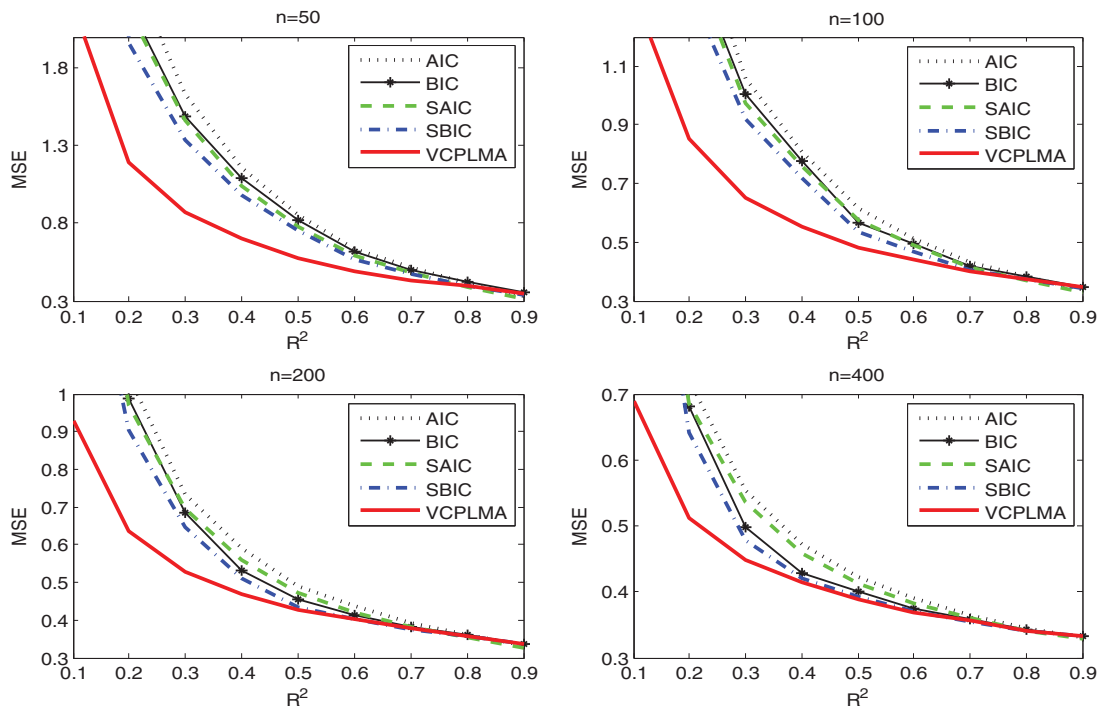


Figure 3. MSE comparisons under Design 6.

Specifically,

$$NMSPE_{\text{mean}}^{(d)} = \frac{1}{D} \sum_{d=1}^D \left(MSPE^{(d)} / R_0^{(d)} \right) \quad (19)$$

and

$$NMSPE_{\text{median}}^{(d)} = \text{median}_{d=1,2,\dots,D} \left(MSPE^{(d)} / R_0^{(d)} \right), \quad (20)$$

where $MSPE^{(d)} = \frac{1}{n_1} \sum_{i=n_0+1}^n (y_i^{(d)} - \hat{\mu}_i^{(d)})^2$, $R_0^{(d)} = \min_{m=1,2,\dots,M} \frac{1}{n_1} \sum_{i=n_0+1}^n (y_i^{(d)} - \hat{\mu}_{(m)i}^{(d)})^2$, $\hat{\mu}_i^{(d)}$ is the predicted value of $y_i^{(d)}$ obtained by a given method in the d th trial, and $\hat{\mu}_{(m)i}^{(d)}$ indicates that the prediction is based on the m th model. They are shown in Table 2, which also reports the optimal rate of each method, defined as the proportion of times in which the method results in the smallest NMSPE across the D replication trials.

The results show that the VCPLMA method is the overwhelming favorite of all methods no matter the performance yardstick, a finding we consider remarkable. The superiority of the VCPLMA method over other strategies is most apparent in terms of optimal rate, for which the VCPLMA estimator always attains a score of over 50%, meaning that in over half of the trials, the VCPLMA method yields the smallest NMSPE among the five estimators. The SAIC estimator frequently yields an optimal rate that is a distant second to the VCPLMA estimator but best among the four remaining estimators. In terms of the mean and median of NMSPE, the SBIC estimator has an edge over the SAIC estimator, but it is never found to be a better alternative than the VCPLMA estimator. The SAIC and SBIC estimators both improve over their model selection counterparts in terms of mean, median, and optimal rate.

Table 3 reports the Diebold and Mariano (1995) test results for the differences in MSPE. The test statistics and p -values presented in Columns 4, 7, 9, and 10 of the table show that the differences in MSPE between the VCPLMA estimator and the other four estimators are all statistically significant. The test results shown in Columns 2, 3, 5, and 6 indicate the same about the differences between each of the SAIC and SBIC estimators and the two selection-based estimators. However, the same cannot be said about the difference in performance between the SAIC

Table 2. Normalized mean squared prediction errors (NMSPE) of five methods ($D = 500$).

n_0	Method	AIC	BIC	SAIC	SBIC	VCPLMA
600	Mean	1.019	1.017	1.012	1.012	1.007
	Median	1.016	1.014	1.010	1.010	1.006
	Optimal rate	0.072	0.028	0.178	0.152	0.570
700	Mean	1.017	1.016	1.012	1.011	1.007
	Median	1.014	1.013	1.009	1.009	1.006
	Optimal rate	0.098	0.028	0.190	0.136	0.548
800	Mean	1.017	1.016	1.013	1.012	1.008
	Median	1.012	1.013	1.010	1.009	1.007
	Optimal rate	0.112	0.032	0.180	0.176	0.500
900	Mean	1.018	1.018	1.014	1.014	1.010
	Median	1.013	1.015	1.011	1.012	1.009
	Optimal rate	0.152	0.022	0.188	0.134	0.504
1000	Mean	1.021	1.021	1.017	1.017	1.012
	Median	1.017	1.018	1.015	1.014	1.012
	Optimal rate	0.126	0.042	0.162	0.116	0.554
1100	Mean	1.026	1.024	1.021	1.021	1.017
	Median	1.022	1.021	1.018	1.019	1.016
	Optimal rate	0.110	0.068	0.236	0.140	0.446
1200	Mean	1.039	1.037	1.033	1.035	1.032
	Median	1.035	1.030	1.030	1.031	1.029
	Optimal rate	0.130	0.110	0.284	0.100	0.376

Table 3. Diebold–Mariano statistics: Mean squared prediction errors (MSPE) ($D = 500$).

n_0		AIC BIC	AIC SAIC	AIC SBIC	AIC VCPLMA	BIC SAIC	BIC SBIC	BIC VCPLMA	SAIC SBIC	SAIC VCPLMA	SBIC VCPLMA
600	DM	1.905	15.715	9.876	17.418	4.667	8.409	11.514	1.699	11.550	12.081
	p-Value	0.057	0.000	0.000	0.000	0.000	0.000	0.000	0.089	0.000	0.000
700	DM	1.761	15.545	9.282	16.879	5.356	12.455	13.303	1.351	11.227	10.801
	p-Value	0.078	0.000	0.000	0.000	0.000	0.000	0.000	0.177	0.000	0.000
800	DM	2.429	14.543	9.183	15.401	5.930	17.205	18.761	1.891	10.158	11.649
	p-Value	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.059	0.000	0.000
900	DM	0.962	12.637	6.576	12.750	6.459	15.034	18.427	0.175	9.060	12.632
	p-Value	0.336	0.000	0.000	0.000	0.000	0.000	0.000	0.861	0.000	0.000
1000	DM	0.437	11.034	5.406	12.460	6.379	14.463	19.317	-0.708	9.680	14.572
	p-Value	0.662	0.000	0.000	0.000	0.000	0.000	0.000	0.479	0.000	0.000
1100	DM	2.610	13.443	6.203	11.628	3.842	9.724	13.118	0.030	7.582	10.780
	p-Value	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.976	0.000	0.000
1200	DM	1.940	11.003	4.328	7.703	3.020	3.330	6.024	-1.870	2.843	6.763
	p-Value	0.052	0.000	0.000	0.000	0.003	0.001	0.000	0.061	0.004	0.000

NOTE: A positive Diebold–Mariano statistic indicates that the estimator in the numerator produces a larger MSPE than the estimator in the denominator.

and SBIC estimators, and between the AIC and BIC estimators, as shown in Columns 1 and 8.

7. Concluding Remarks

In the context of the varying-coefficient partially linear model, we have demonstrated a Mallows-type VCPLMA estimator that possesses a large sample justification and has excellent finite sample properties relative to traditional competing model selection and averaging methods. All things considered, the results suggest that the VCPLMA estimator represents a credible alternative that deserves further attention from both theoretical and applied statisticians.

There are a number of ways the present approach can be extended that may result in an even more effective estimator. For example, model screening could be introduced into the analysis, and judging from existing results (e.g., Yuan and Yang 2005), it is conceivable that removing the poorest models before averaging can contribute to greater estimation and predictive efficiency. Also, although we allow the dimension parameters r_m and k_m to increase with n , the sample size, they are not allowed to be greater than n , and their rates of increase are constrained by Condition (C.8). The development of an optimal model averaging method for high-dimensional VCPLM is an intriguing possible extension of the current analysis. There is also room for an extension of the present approach to the generalized varying-coefficient partially linear model (Li and Liang 2008; Lam and Fan 2008) that permits a discrete response variable, and more versatile link functions for error distributions. These remain for future research.

Appendix: Sketches of the Proofs of Results

This appendix contains sketches of the proofs of Theorems 1 and 2. Detailed proofs can be found in the online supplemental file.

A.1 Preliminary Results

The proofs of Theorems 1 and 2 require the following lemmas. The proofs of the lemmas are given in the online supplemental file.

Lemma A.1. Let Conditions (C.3)–(C.5) hold. Then for all $m = 1, 2, \dots, M$ and $t \in \Delta$, we have

$$\frac{1}{n} D_t^T W_{(m)t} D_t = \begin{pmatrix} f(t) + O_{up}(h_m^2) & \mu_2(K)h_m f'(t) + O_{up}(h_m^2) \\ \mu_2(K)h_m f'(t) + O_{up}(h_m^2) & \mu_2(K)f(t) + O_{up}(h_m^2) \end{pmatrix} \otimes C_z$$

and

$$\left\{ \frac{1}{n} D_t^T W_{(m)t} D_t \right\}^{-1} = \begin{pmatrix} f^{-1}(t) + O_{up}(h_m^2) & -h_m f'(t) f^{-2}(t) + O_{up}(h_m^2) \\ -h_m f'(t) f^{-2}(t) + O_{up}(h_m^2) & \mu_2^{-1}(K) f^{-1}(t) + O_{up}(h_m^2) \end{pmatrix} \otimes C_z^{-1},$$

where $\mu_2(K) = \int_{v \in \text{supp}(K)} K(v)v^2 dv$, and if a function $g(t) = O_{up}(b_m^2)$, then $g(t)/b_m^2$ is bounded in probability uniformly for any t within the interior of Δ .

Lemma A.2. Assume that Conditions (C.3)–(C.5) are satisfied. Then we have

$$\max_{1 \leq m \leq M} |tr(A_{(m)})| = O_p(h^{-1}\tilde{r}),$$

$$\max_{1 \leq m \leq M} \bar{\lambda}(A_{(m)}) = O_p(\tilde{r}^{1/2}),$$

and

$$\max_{1 \leq m \leq M} \bar{\lambda}(P_{(m)}) = O_p(\tilde{r}^{1/2}).$$

A.2 Proof of Theorem 1

Note that $C_n(w)$ can be written as

$$\begin{aligned} C_n(w) &= \|Y - \hat{\mu}(w)\|^2 + 2tr(P(w)\Omega) \\ &= L_n(w) + \|\epsilon\|^2 - 2\epsilon^T(P(w) - I_n)\mu \\ &\quad - 2\{\epsilon^T P(w)\epsilon - tr(P(w)\Omega)\}, \end{aligned}$$

where $\|\epsilon\|^2$ is independent of w . Hence to prove [Theorem 1](#), we need only to verify that

$$\sup_{w \in \mathcal{H}_n} |\epsilon^\top (P(w) - I_n)\mu| / R_n(w) = o_p(1), \tag{A.1}$$

$$\sup_{w \in \mathcal{H}_n} |\epsilon^\top P(w)\epsilon - \text{tr}(P(w)\Omega)| / R_n(w) = o_p(1), \tag{A.2}$$

and

$$\sup_{w \in \mathcal{H}_n} |L_n(w)/R_n(w) - 1| = o_p(1). \tag{A.3}$$

Note that

$$\begin{aligned} & \sup_{w \in \mathcal{H}_n} |L_n(w)/R_n(w) - 1| \\ &= \sup_{w \in \mathcal{H}_n} |2\epsilon^\top P^\top(w)(P(w) - I_n)\mu \\ & \quad + \|P(w)\epsilon\|^2 - \text{tr}(P^\top(w)P(w)\Omega)| / R_n(w). \end{aligned}$$

It suffices to show that

$$\sup_{w \in \mathcal{H}_n} |\epsilon^\top P^\top(w)(P(w) - I_n)\mu| / R_n(w) = o_p(1), \tag{A.4}$$

and

$$\sup_{w \in \mathcal{H}_n} \left| \|P(w)\epsilon\|^2 - \text{tr}(P^\top(w)P(w)\Omega) \right| / R_n(w) = o_p(1). \tag{A.5}$$

As the variables X, Z, Ψ are random, we first prove, for any $\delta > 0$, that

$$P \left(\sup_{w \in \mathcal{H}_n} |\epsilon^\top (P(w) - I_n)\mu| / R_n(w) > \delta \mid X, Z, \Psi \right) = o_p(1),$$

then we have $P(\sup_{w \in \mathcal{H}_n} |\epsilon^\top (P(w) - I_n)\mu| / R_n(w) > \delta) \rightarrow 0$, which is Equation (A.1). By similar steps, we can prove (A.2), (A.4), and (A.5). This proves [Theorem 1](#). A detailed proof of [Theorem 1](#) is given in the online supplemental file.

A.3 Proof of Theorem 2

When Ω is replaced by $\hat{\Omega}$, $C_n(w)$ is correspondingly changed to $\hat{C}_n(w) = C_n(w) + 2\{\text{tr}(P(w)\hat{\Omega}) - \text{tr}(P(w)\Omega)\}$. From the result of [Theorem 1](#), to prove [Theorem 2](#), it suffices to prove that

$$\sup_{w \in \mathcal{H}_n} \left| \text{tr}(P(w)\hat{\Omega}) - \text{tr}(P(w)\Omega) \right| / R_n(w) = o_p(1). \tag{A.6}$$

Let $Q_{(m)} = \text{diag}(\rho_{11}^{(m)}, \dots, \rho_{nn}^{(m)})$ and $Q(w) = \sum_{m=1}^M w_m Q_{(m)}$. To prove (A.6), we decompose the left-hand side of (A.6) into five parts as follows:

$$\begin{aligned} & \sup_{w \in \mathcal{H}_n} \left| \text{tr}(P(w)\hat{\Omega}) - \text{tr}(P(w)\Omega) \right| / R_n(w) \\ &= \sup_{w \in \mathcal{H}_n} \left| (Y - P_{(M^*)}Y)^\top Q(w)(Y - P_{(M^*)}Y) \right. \\ & \quad \left. - \text{tr}(Q(w)\Omega) \right| / R_n(w) \\ &= \sup_{w \in \mathcal{H}_n} \left| (\epsilon + \mu)^\top (I_n - P_{(M^*)})^\top Q(w)(I_n - P_{(M^*)})(\epsilon + \mu) \right. \\ & \quad \left. - \text{tr}(Q(w)\Omega) \right| / R_n(w) \\ &\leq \sup_{w \in \mathcal{H}_n} \left| \epsilon^\top (I_n - P_{(M^*)})^\top Q(w)(I_n - P_{(M^*)})\epsilon \right. \\ & \quad \left. - \text{tr}\{(I_n - P_{(M^*)})^\top Q(w)(I_n - P_{(M^*)})\Omega\} \right| / R_n(w) \\ & \quad + 2 \sup_{w \in \mathcal{H}_n} \left| \epsilon^\top (I_n - P_{(M^*)})^\top Q(w)(I_n - P_{(M^*)})\mu \right| / R_n(w) \\ & \quad + \sup_{w \in \mathcal{H}_n} \left| \mu^\top (I_n - P_{(M^*)})^\top Q(w)(I_n - P_{(M^*)})\mu \right| / R_n(w) \\ & \quad + \sup_{w \in \mathcal{H}_n} \left| \text{tr}\{P_{(M^*)}^\top Q(w)P_{(M^*)}\Omega\} \right| / R_n(w) \\ & \quad + 2 \sup_{w \in \mathcal{H}_n} \left| \text{tr}\{P_{(M^*)}^\top Q(w)\Omega\} \right| / R_n(w) \\ & \equiv \Xi_1 + \Xi_2 + \Xi_3 + \Xi_4 + \Xi_5. \end{aligned}$$

Now, define $\rho = \max_{1 \leq m \leq M} \max_{1 \leq i \leq n} |\rho_{ii}^{(m)}|$. From [Lemma A.2](#) and [Condition \(C.6\)](#), we have $\rho = O_p(n^{-1}\tilde{r}^{1/2}\tilde{k} + n^{-1}h^{-1}\tilde{r})$. It follows from [Equation \(10\)](#) and [Condition \(C.2\)](#) that $\xi_n^{-1} = o_p(1)$, $M\xi_n^{-2G}\tilde{r}^G = o_p(1)$, and $\xi_n^{-2}\tilde{r}\|P_{(M^*)}\mu - \mu\|^2 = o_p(1)$. Using [Lemma A.2](#), [Equation \(10\)](#), [Conditions \(C.2\), \(C.6\), and \(C.8\)](#), Chebyshev's inequality, and [Theorem 2](#) of [Whittle \(1960\)](#), we can obtain, for any $\delta > 0$, that $P(\Xi_1 > \delta \mid X, Z, \Psi) = o_p(1)$. Then we have $P(\Xi_1 > \delta) = o(1)$, which implies that $\Xi_1 = o_p(1)$. We can also prove that each of Ξ_2, Ξ_3 and $\Xi_4 + \Xi_5$ is equal to $o_p(1)$. This proves [Theorem 2](#). A detailed proof of [Theorem 2](#) is given in the online supplemental file.

Supplementary Material

The online supplemental file contains the proofs of [Lemmas A.1](#) and [A.2](#) and detailed proofs of [Theorems 1](#) and [2](#).

Acknowledgment

The authors thank the editor, associate editor, and two referees for insightful comments that have helped improve the quality of the article. The usual disclaimer applies.

Funding

Wan's work was supported by a Strategic Grant from the City University of Hong Kong (Grant no. 7004786). Zhang's and Zou's work was supported by the following funding bodies: National Natural Science Foundation of China (Grant nos. 71522004 (Zhang), 11471324 (Zhang), 71631008 (Zhang), 11331011 (Zou), and 11529101 (Zou)) and the Ministry of Science and Technology of China (Grant no. 2016YFB0502301 (Zou)).

References

Ahmad, I., Leelahanon, S., and Li, Q. (2005), "Efficient Estimation of a Semiparametric Partially Linear Varying Coefficient Model," *Annals of Statistics*, 33, 258–283. [882]
 Ando, T., and Li, K.-C. (2014), "A Model-Averaging Approach for High-Dimensional Regression," *Journal of the American Statistical Association*, 109, 254–265. [885]
 Andrews, D. W. K. (1991), "Asymptotic Optimality of Generalized C_L , Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377. [886]

- Box, G. (1976), "Science and Statistics," *Journal of the American Statistical Association*, 71, 791–799. [884]
- Box, G., and Draper, N. (1987), *Empirical Model Building and Response Surfaces*, New York: Wiley. [884]
- Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [883,886]
- Cleveland, W. S., Grosse, E., and Shyu, W. M. (1991), *Local Regression Models*, Pacific Grove, CA: Wadsworth and Brooks/Cole. [882]
- Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [889]
- 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. [882]
- Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057. [882,884]
- Fan, J., and Jiang, J. (2007), "Nonparametric Inference with Generalized Likelihood Ratio Tests," *Test*, 16, 409–444. [882]
- Fan, J., Lin, H., and Zhou, Y. (2006), "Local Partial-Likelihood Estimation for Lifetime Data," *Annals of Statistics*, 34, 290–325. [887]
- Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *Annals of Statistics*, 29, 153–193. [882]
- Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [883,884,885]
- Hansen, B. E., and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [884,886]
- Härdle, W., Liang, H., and Gao, J. (2000), *Partially Linear Models*, Heidelberg: Physica-Verlag. [882]
- Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society, Series B*, 55, 757–796. [882]
- Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [883]
- (2006), "Focused Information Criteria and Model Averaging for the Cox's Hazard Regression Model," *Journal of the American Statistical Association*, 101, 1449–1464. [883,887]
- Lam, C., and Fan, J. (2008), "Profile-Kernel Likelihood Inference with Diverging Number of Parameters," *Annals of Statistics*, 36, 2232–2260. [890]
- Li, Q., Huang, C. J., Li, D., and Fu, T. T. (2002), "Semiparametric Smooth Coefficient Models," *Journal of Business & Economic Statistics*, 20, 412–422. [882]
- Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Modeling," *Annals of Statistics*, 36, 261–286. [890]
- 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. [886]
- Liu, Q., and Okui, R. (2013), "Heteroskedasticity-Robust Cp Model Averaging," *Econometrics Journal*, 16, 463–472. [883,884,885,886]
- Lu, X., and Su, L. (2015), "Jackknife Model Averaging for Quantile Regressions," *Journal of Econometrics*, 188, 40–58. [883]
- Morris, C. N., Norton, E. C., and Zhou, X. H. (1994), "Parametric Duration Analysis of Nursing Home Usage," in *Case Studies in Biometry*, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, J. Greenhouse, pp. 231–248, New York: Wiley. [887]
- Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society, Series B*, 50, 413–436. [882]
- Wan, A. T. K., Zhang, X., and Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283. [884,885,886]
- Wang, H., Zou, G., and Wan, A. T. K. (2012), "Model Averaging for Varying-Coefficient Partially Linear Measurement Error Models," *Electronic Journal of Statistics*, 6, 1017–1039. [883,885]
- Wang, H. J., Zhu, Z., and Zhou, J. (2009), "Quantile Regression in Partially Linear Varying Coefficient Models," *Annals of Statistics*, 37, 3841–3866. [882]
- Whittle, P. (1960), "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability & its Applications*, 5, 331–335. [891]
- Xia, Y., Zhang, W., and Tong, H. (2004), "Efficient Estimation for Semivarying-Coefficient Models," *Biometrika*, 91, 661–681. [882]
- Xie, S., Wan, A. T. K., and Zhou, Y. (2015), "Quantile Regression Methods with Varying-Coefficient Models for Censored Data," *Computational Statistics & Data Analysis*, 88, 154–172. [887]
- 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. [884]
- Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214. [890]
- Zhang, W., Lee, S. Y., and Song, X. (2002), "Local Polynomial Fitting in Semivarying Coefficient Model," *Journal of Multivariate Analysis*, 82, 166–188. [882]
- 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 Nonzero Threshold," *Journal of Business & Economic Statistics*, 30, 132–142. [883,887]
- Zhang, X., and Wang, W. (2018), "Optimal Model Averaging Estimation for Partially Linear Models," *Statistica Sinica*, preprint. [883]
- Zhao, P., and Xue, L. (2010), "Variable Selection for Semiparametric Varying Coefficient Partially Linear Errors-in-Variables Models," *Journal of Multivariate Analysis*, 101, 1872–1883. [882]