Statistica Sinica Preprint No: SS-2017-0034.R2	
Title	OPTIMAL MODEL AVERAGING OF VARYING COEFFICIENT MODELS
Manuscript ID	SS-2017-0034.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0034
Complete List of Authors	CONG LI QI LI JEFFREY S. RACINE and DAIQIANG ZHANG
Corresponding Author	Jeffrey S. Racine
E-mail	racinej@mcmaster.ca

Statistica Sinica

Optimal Model Averaging of Varying Coefficient Models

Cong Li

School of Economics, Shanghai University of Finance and Economics,

777 Guoding Road, Shanghai, 200433, PR China, li.cong@mail.shufe.edu.cn.

Qi Li

ISEM, Capital University of Economics and Business,

Beijing, 100070, PR China; Department of Economics, Texas A&M University,

College Station, TX 77843, USA, qi-li@tamu.edu.

Jeffrey S. Racine

Department of Economics and Graduate Program in Statistics,

McMaster University, racinej@mcmaster.ca;

Department of Economics and Finance, La Trobe University;

Info-Metrics Institute, American University;

Rimini Center for Economic Analysis; Center for Research

in Econometric Analysis of Time Series (CREATES), Aarhus University.

Daiqiang Zhang

Department of Economics, University at Albany,

SUNY, Albany, NY 12222, USA, dzhang6@albany.edu.

Abstract: **We consider the problem of model averaging over a set of semiparametric varying coefficient models where the varying coefficients can be functions of continuous and categorical variables. We propose a Mallows model averaging procedure that is capable of delivering model averaging estimators with solid finite-sample performance. Theoretical underpinnings are provided, finite-sample performance is assessed via Monte Carlo simulation, and an illustrative application is presented. The approach is very simple to implement in practice and R code is provided in an supplementary material.**

1991 Mathematics Subject Classification: C14 Semiparametric and Nonparametric Methods

Key words and phrases: Kernel Smoothing, Semiparametric, Candidate Models.

1. Introduction

Practitioners who wish to tackle *model uncertainty* have a variety of approaches at their disposal. The most promising involve *model selection* and *model averaging*. *Model selection* proceeds from the premise that all models are, at best, approximations and involves selecting one model from

among a set of candidate models. It is understood that, in practice, it is unlikely that the *true* model is among the set of candidate models, hence the model selected is the least misspecified among the set of models considered, in some known statistical sense. In essence, the practitioner who adopts model selection applies weight 1 to one candidate model and weight 0 to all others using a *selection criterion*. Model selection has a long history, and a variety of methods have been proposed, each based on distinct estimation criteria. These include Akaike's *An Information Criterion* (AIC; Akaike (1970), Akaike (1973)), Mallows' C_p (Mallows (1973)), the *Bayesian Information Criterion* (BIC; Schwarz (1978)), *delete-one cross-validation* (Stone (1974)), *generalized cross-validation* (Craven and Wahba (1979)), and the *Focused Information Criterion* (FIC) (Claeskens and Hjort (2003)), to name but a few.

Model averaging, on the other hand, produces a model that is a *weighted average* defined over a set of candidate models for which the weights are chosen by a statistical procedure having known properties, an *averaging criterion*. There is a longstanding literature on Bayesian model averaging; see Hoeting, Madigan, Raftery, and Volinsky (1999) for a comprehensive review. There is also a rapidly-growing literature on frequentist methods for model averaging, including Buckland, Burnhamn, and Augustin

(1997),Hansen (2007), Wan, Zhang, and Zou (2010), Hansen and Racine (2012), Zhang and Wang (2015), Zhang, Zou, and Carroll (2015) and Zhang, Yu, Zou, and Liang (2016), among others.

Practitioners who adopt the model averaging approach often construct a weighted average defined over a set of *parametric* candidates. An alternative approach, one that we consider here, is to instead construct a weighted average defined over a set of more flexible *semiparametric* candidates. From a practical perspective, one might hope that by using more flexible estimators for the set of candidate models perhaps fewer candidate models might be needed, or that perhaps the approximation capabilities of the resulting model might be improved. Though one might be tempted to perhaps average over fully nonparametric models, such models suffer from the so-called *curse of dimensionality* and are restricted to only a few predictors at most. Semiparametric models strike a balance between flexibility and efficiency thereby attenuating the curse of dimensionality. Furthermore, being semiparametric in nature, one can easily incorporate prior parametric information if it exists. Zhang and Wang (2015) is the first to consider averaging over Robinson's (1988) semiparametric partially linear model. Our approach involves averaging over the so-called varying coefficient specification; see Beran and Hall (1992), Hastie and Tibshirani (1993), Cai, Fan, and

Yao (2000), Li, Huang, Li, and Fu (2002) and the references therein. The varying coefficient specification is particularly appealing in this context, in part because a range of models turns out to be special cases including a fully nonparametric model and Robinson's (1988) partially linear model, by way of illustration. Our approach adopts Mallows' C_p criterion (Mallows (1973)) for selecting the averaging weights, and allows for the coefficients in the varying coefficient candidate models to be functions of either continuous data types, categorical data types, or a mix of both.

Our theoretical results (based on the Mallows criterion) apply both to nested and non-nested regression models, and allow for heterogeneous errors. Hansen (2014) examines the asymptotic risk of nested least-squares averaging estimators based on minimizing a generalized Mallows criterion in a linear model with heteroskedasticity. Liu, Okui, and Yoshimura (2016) adopt the Mallows criterion to choose the weight vector in the model averaging estimator for linear regression models with heteroskedastic errors. By averaging over semiparametric specifications we generalize existing approaches and provide practitioners with a straightforward and powerful approach to handling model uncertainty.

The rest of this paper proceeds as follows. Section 2 presents the varying coefficient specification defined over mixed datatypes, Mallows-driven

weight choice, and asymptotic optimality of the proposed approach. Section 3 examines the finite-sample performance of the proposed approach relative to alternative model averaging estimators and model selection estimators, while Section 4 considers an illustrative example and a comparison of hold-out data performance for a range of averaging and selection criteria. Section 5 presents some brief concluding remarks. Proofs of the main theorems are provided in Supplementary Material 1, while R code can be found in Supplementary Material 2.

2. Model Averaging Estimation

2.1 Model and estimators

We consider a varying coefficient model

$$Y_i = \mu_i + \epsilon_i = \sum_{j=1}^{\infty} X_{ij}\beta_j(Z_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $X_i = (X_{i1}, X_{i2}, \cdots)'$ is a countably infinite random vector, $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ is a $q \times 1$ random vector, $\beta(Z_i) = (\beta_1(Z_i), \beta_2(Z_i), \ldots)'$ is a countably infinite unknown vector function, $\mu_i = X_i'\beta(Z_i)$, the idiosyncratic error term ϵ_i is possibly conditionally heteroscedastic satisfying $E(\epsilon_i|X_i, Z_i) = 0$ and $E(\epsilon_i^2|X_i, Z_i) = \sigma_i^2$. The observations $(X_i, Z_i, Y_i)_{i=1}^n$ are independent

across i.

Our goal is to estimate μ_i for the purposes of prediction, the focus of the literature on model averaging estimation; see Hansen (2007) and Lu and Su (2015) by way of illustration. To this end, we use S_n candidate varying coefficient models to approximate (2.1), where the number of models, S_n, is allowed to diverge to infinity as $n \to \infty$. The s_{th} candidate model is

$$Y_i = X'_{i,(s)}\beta_{(s)}(Z_{i,(s)}) + b_{i,(s)} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where $X'_{i,(s)}$ is a p_s-dimensional subset of X_i, $Z_{i,(s)}$ is a q_s-dimensional ($1 \leq q_s \leq q$) subset of Z_i, $\beta_{(s)}(Z_{i,(s)})$ is the corresponding $p_s \times 1$ unknown function, and $b_{i,(s)} = \mu_i - X'_{i,(s)}\beta_{(s)}(Z_{i,(s)})$ represents the approximation error in the s_{th} model.

To provide an optimal weighting scheme, we first need to estimate each candidate model. Premultiplying (2.1) by $X_{i,(s)}$ and taking $E(\cdot|Z_{i,(s)} = z_{(s)})$ leads to $E[X_{i,(s)}Y_i|Z_{i,(s)} = z_{(s)}] = E[X_{i,(s)}X'_{i,(s)}]\beta_{(s)}(z_{(s)})$, yielding

$$\beta_{(s)}(z_{(s)}) = [E(X_{i,(s)}X'_{i,(s)}|z_{(s)})]^{-1}E[X_{i,(s)}Y_i|z_{(s)}]. \tag{2.3}$$

Let $K_{(s)}\left(\frac{Z_{j,(s)}-z_{(s)}}{h_{(s)}}\right) = k_1\left(\frac{Z_{j,(s),1}-z_{(s),1}}{h_{(s),1}}\right) \times \ldots \times k_{q_s}\left(\frac{Z_{j,(s),q_s}-z_{(s),q_s}}{h_{(s),q_s}}\right)$ denote a *product kernel function*, where $k(\cdot)$ is a univariate kernel function and

$h_{(s),r}$ is a scalar bandwidth for $r = 1, \ldots, q_s$. When the data consist of a mix of categorical and continuous datatypes, one can replace the above kernel function by the generalized kernel function that smooths both the continuous and the discrete covariates; see Hall, Racine, and Li (2004) for details, and also Hall, Li, and Racine (2007), and Hall and Racine (2015) for related extensions. Then (2.3) suggests a local constant least-squares estimator,

$$\widehat{\beta}_{(s)}(z_{(s)}) = \left[\sum_{j=1}^{n} X_{j,(s)} X'_{j,(s)} K_{(s)}\left(\frac{Z_{j,(s)} - z_{(s)}}{h_{(s)}}\right)\right]^{-1} \sum_{j=1}^{n} X_{j,(s)} Y_j K_{(s)}\left(\frac{Z_{j,(s)} - z_{(s)}}{h_{(s)}}\right). \tag{2.4}$$

Letting $X_{(s)} = (X_{1,(s)}, \ldots, X_{n,(s)})'$, $Z_{(s)} = (Z_{1,(s)}, \ldots, Z_{n,(s)})'$, $Y = (Y_1, \ldots, Y_n)'$, and $\mathcal{K}_{[z_{(s)}]}$ be an $n \times n$ diagonal matrix with jth diagonal element $K_{(s)}\left(\frac{Z_{j,(s)} - z_{(s)}}{h_{(s)}}\right)$, we can rewrite (2.4) as

$$\widehat{\beta}_{(s)}(z_{(s)}) = \left(X'_{(s)} \mathcal{K}_{[z_{(s)}]} X_{(s)}\right)^{-1} X'_{(s)} \mathcal{K}_{[z_{(s)}]} Y. \tag{2.5}$$

Then, we can estimate $\mu_{i,(s)}$ by

$$\widehat{\mu}_{i,(s)} = X'_{i,(s)} \widehat{\beta}_{(s)}(Z_{i,(s)}) = X'_{i,(s)} \left(X'_{(s)} \mathcal{K}_{[Z_{i,(s)}]} X_{(s)}\right)^{-1} X'_{(s)} \mathcal{K}_{[Z_{i,(s)}]} Y, \tag{2.6}$$

and rewrite it in matrix notation as $\widehat{\mu}_{(s)} = P_{(s)} Y$, where $P_{(s)}$ is a square ma-

trix of dimension $n \times n$ with ith row $X'_{i,(s)} \left(X'_{(s)} \mathcal{K}_{[Z_{i,(s)}]} X_{(s)} \right)^{-1} X'_{(s)} \mathcal{K}_{[Z_{i,(s)}]}$, and $\widehat{\mu}_{(s)} = (\widehat{\mu}_{1,(s)}, \ldots, \widehat{\mu}_{n,(s)})'$. Let the weight vector $w = (w_1, \ldots, w_{S_n})^T$ belong to the set $\mathcal{W} = \{w \in [0,1]^{S_n} : \sum_{s=1}^{S_n} w_s = 1\}$, and let $P(w) = \sum_{s=1}^{S_n} w_s P_{(s)}$. Then, the model averaging estimator of μ is specified as

$$\widehat{\mu}(w) = \sum_{s=1}^{S_n} w_s \widehat{\mu}_{(s)} = P(w)Y. \tag{2.7}$$

2.2 Weight Choice Criterion and Asymptotic Optimality

Until now, the weight vector in $\widehat{\mu}(w)$ was left unspecified. Motivated by the Mallows criterion for model averaging estimators (e.g. Hansen (2007), we now outline how we choose this weight vector. Let $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Define the predictive squared loss by

$$L_n(w) = n^{-1} \|\widehat{\mu}(w) - \mu\|^2, \tag{2.8}$$

and the conditional expected loss by

$$R_n(w) = E[L_n(w)|X, Z] = n^{-1} \|P(w)\mu - \mu\|^2 + n^{-1} \,\text{trace}[\Omega P(w)' P(w)]. \tag{2.9}$$

Let the Mallows-type criterion function be

$$C_n(w) = n^{-1}\|P(w)Y - Y\|^2 + 2n^{-1}\,\mathrm{trace}[P(w)\Omega]. \tag{2.10}$$

It is easy to show that

$$R_n(w) = E[C_n(w)|X, Z] - n^{-1}\,\mathrm{trace}(\Omega),$$

which suggests that, for the optimal choice of w in the sense of minimizing $R_n(w)$, we can minimize $C_n(w)$ to choose w by noting that $n^{-1}\,\mathrm{trace}(\Omega)$ does not depend on w. Assuming that Ω is known, the optimal weight choice is given by

$$\widehat{w} = \mathrm{argmin}_{w\in\mathcal{W}}\, C_n(w), \tag{2.11}$$

which implies that the optimal model averaging estimator of μ is $\widehat{\mu}(\widehat{w}) = P(\widehat{w})Y$, and we refer to $\widehat{\mu}(\widehat{w})$ as a *Mallows model average of varying coefficient models*. In order to provide regularity conditions for the optimal choice of the weight vector, we need to introduce some notation. Let $\xi_n = \inf_{w\in\mathcal{W}} nR_n(w)$, and let w_s^o be an $S_n \times 1$ vector in which the sth element is one and all others are zeros. Here are the conditions required for the

asymptotic optimality of $\widehat{w}$ as defined in (2.11). Given the randomness of X and Z, the following conditions and related proofs presented elsewhere in the paper are to hold almost surely; For brevity, we omit the phrase "almost surely". Let $\overline{p} = \max_{1 \leq s \leq S_n} p_s$. For some integer $N \geq 1$,

$$\max_i E(\epsilon_i^{4N} | X_i, Z_i) < \infty, \tag{2.12}$$

$$S_n \overline{p}^{4N} \xi_n^{-2N} \sum_{s=1}^{S_n} [nR_n(w_s^o)]^N \to 0, \tag{2.13}$$

$$\sup_{s \in \{1,\ldots,S_n\}} \max_i \sum_{j=1}^n |P_{(s),ij}| = O(\overline{p}^2) \quad \text{and} \quad \sup_{s \in \{1,\ldots,S_n\}} \max_j \sum_{i=1}^n |P_{(s),ij}| = O(\overline{p}^2). \tag{2.14}$$

The first two conditions are commonplace in the literature on model averaging estimation (e.g., Hansen (2007); Hansen and Racine (2012); Wan, Zhang, and Zou (2010); Ando and Li (2014)). Condition (2.13) requires that $\xi_n \to \infty$, implying that there is no finite approximating model whose bias is zero. This condition also constrains the rates at which S_n and $nR_n(w_s^o)$ approach ∞.

Condition (2.14) is a somewhat high level assumption. It implicitly

imposes some conditions on the smoothing parameters, such as $h_{(s),j} \to 0$ for all $j = 1, \ldots, q_s$ and $nH_{(s)} \to \infty$ for all $s = 1, \ldots, S_n$, where $H_{(s)} = h_{(s),1} \times \cdots \times h_{(s),q_s}$. As shown in Supplementary Material 1, we provide sufficient regularity conditions on the smoothing parameters and the boundedness and full rank of X needed to obtain (2.14). Analogously, Speckman (1988) uses the kernel smoothing to define the weighting matrix and imposes a weaker bound condition $O(1)$. We conjecture that it may be possible to relax the condition $\max_i \sum_{j=1}^n |P_{(s),ij}| = O(\overline{p}^2)$ to $\max_i \sum_{j=1}^n |P_{(s),ij}| = O(1)$, as used in Speckman (1988) and Zhang and Wang (2015). We leave the verification of this conjecture for future investigation. In practice, one can select the bandwidth for each candidate model by the typical least-squares cross-validation method, and in our simulations we use the cross-validation method that allows for different bandwidths across covariates, and across different candidate models.

Theorem 1. *Under conditions (2.12)-(2.14),*

$$\frac{L_n(\widehat{w})}{\inf_{w \in \mathcal{W}} L_n(w)} \to 1$$

in probability as $n \to \infty$.

This shows that practitioners can do as well asymptotically as if they

knew the true μ_i, the weight vector $\widehat{w}$ is asymptotically optimal in the sense that the average loss with $\widehat{w}$ is asymptotically equivalent to that using the infeasible optimal weight vector.

So far we have assumed that Ω is known. In practice, Ω will be unknown. To make the Mallows-type criterion (2.10) computationally feasible, we estimate the unknown Ω based on residuals from model averaging estimation by

$$\widehat{\Omega}(w) = \text{diag}(\hat{\epsilon}_1^2(w), \ldots, \hat{\epsilon}_n^2(w)), \tag{2.15}$$

where $\hat{\epsilon}_i(w) = Y_i - \hat{\mu}_i(w)$. Replacing Ω with $\widehat{\Omega}$ in $C_n(w)$, we obtain the feasible criterion

$$\widehat{C}_n(w) = n^{-1}\|P(w)Y - Y\|^2 + 2n^{-1}\,\text{trace}[P(w)\widehat{\Omega}(w)]. \tag{2.16}$$

Correspondingly, the new optimal weights are defined as

$$\widetilde{w} = \text{argmin}_{w\in\mathcal{W}}\,\widehat{C}_n(w). \tag{2.17}$$

We now show that the weight vector $\widetilde{w}$ is still asymptotically optimal. Let $\rho_{ii}^{(s)}$ be the i^{th} diagonal element of $P_{(s)}$. The conditions required for the

asymptotic optimality of $\widetilde{w}$ are as follows.

There exists a constant c such that $|\rho_{ii}^{(s)}| \leq cn^{-1}|\operatorname{trace}(P_{(s)})|$, $\forall s = 1, \ldots, S_n$, (2.18)

$$n^{-1}\overline{p}^2 = O(1). \tag{2.19}$$

Condition (2.18) is commonly used to ensure the asymptotic optimality of cross-validation (e.g., Andrews (1991) and Hansen and Racine (2012)). Condition (2.19), Condition (12) of Wan, Zhang, and Zou (2010), allows the p_s's to increase as $n \to \infty$, but restricts their rate of increase.

Theorem 2. *Under conditions (2.12)-(2.14), (2.18), and (2.19)*

$$\frac{L_n(\widetilde{w})}{\inf_{w \in \mathcal{W}} L_n(w)} \to 1 \tag{2.20}$$

in probability as $n \to \infty$.

It is easy to prove that Theorems 1 and 2 apply to the mixed data setting in which $Z = (Z_c, Z_d)$ with Z_c being a continuous vector and Z_d a discrete vector, because our proofs are valid as long as the model averaging estimator is linear in Y when Z consists of multivariate mixed discrete and

continuous covariates, which continues to be the case.

An alternative strategy for estimating Ω can be based on the largest model indexed by $s^* = \underset{s\in\{1,\cdots,S_n\}}{\operatorname{argmax}}\,(p_s + q_s)$,

$$\widehat{\Omega}_{(s^*)} = \operatorname{diag}(\hat{\epsilon}^2_{s^*,1}, \cdots, \hat{\epsilon}^2_{s^*,n}), \tag{2.21}$$

where $(\hat{\epsilon}_{s^*,1}, \cdots, \hat{\epsilon}_{s^*,n}) = Y - \widehat{\mu}_{(s^*)} = Y - P_{(s^*)}Y$. The idea of using the largest model to estimate the variance parameter or covariance matrix is advocated by Hansen (2007), Liu and Okui (2013), and Zhang and Wang (2015)(If the model with the largest dimension is not uniquely defined because the models with the same dimension can differ in the structure of X_i and Z_i, we adopt the model with the largest dimension of X_i following Zhang and Wang (2015).). The motivation for $\widehat{\Omega}(w)$ in Theorem 2 is to avoid putting too much confidence in a single model while the advantage of $\widehat{\Omega}_{(s^*)}$ is that the computational burden is much less than using $\widehat{\Omega}(w)$ because the estimator of the error covariance matrix $\widehat{\Omega}_{(s^*)}$ does not include the weight vector w, which implies that $\widehat{C}^*_n(w)$ defined in (2.16) below is a lower-order function of w than $\widehat{C}_n(w)$. In particular, using $\widehat{\Omega}_{(s^*)}$ allows us to solve a simple quadratic program that can be done with standard off-the-shelf software. Replacing Ω with $\widehat{\Omega}_{(s^*)}$ in $C_n(w)$, we obtain the feasible

criterion

$$\widehat{C}_n^*(w) = n^{-1}\|P(w)Y - Y\|^2 + 2n^{-1}\,\mathrm{trace}[P(w)\widehat{\Omega}_{(s^*)}].$$

Correspondingly, the new optimal weights are defined as

$$\widetilde{w}_{(s^*)} = \mathrm{argmin}_{w\in\mathcal{W}}\,\widehat{C}_n^*(w).$$

Then, using the definitions of $\rho_{ii}^{(s)}$ and $\overline{p}$ above and the same conditions as in Theorem 2, we can show that the weight vector $\widetilde{w}_{(s^*)}$ is still asymptotically optimal.

Corollary 1. *Under conditions (2.12)-(2.14), (2.18), and (2.19) with the alternative estimators* $\widehat{\Omega}_{(s^*)}$,

$$\frac{L_n(\widetilde{w}_{(s^*)})}{\inf_{w\in\mathcal{W}} L_n(w)} \to 1 \tag{2.22}$$

in probability as $n \to \infty$.

3. Monte Carlo Simulations

In this section we report on the finite-sample performance of the proposed Mallows model averaging ('MMA') method. We simulated data

from an infinite-order varying coefficient regression model of the form $y_i = \sum_{j=1}^{\infty} \theta_j(z_i)x_{ij} + \epsilon_i,\ i = 1, \ldots, n$. The x_{ij} were independent and identically distributed $N(0,1)$ random variates, while z_i was $U[-1,1]$. The heteroskedastic error ϵ_i was distributed $N(0, \sigma^2(z_i))$, where $\sigma(z_i) = \sigma|z_i|\sqrt{3}$ and independent of the x_{ij}.

The parameters were determined by the rule $\theta_j(z_i) = \sqrt{2\alpha}j^{-\alpha-1/2}\exp(z_i)$. The sample size was $n = 50, 100, 200$, and 400. The parameter α was 0.10, 0.25, and 0.50. Larger values of α imply that the coefficients $\theta_j(z)$ decline more quickly with j. The number of models M_n was determined by the rule $M_n = 3n^{1/3}$ (so $M_n = 11, 14, 18$, and 22 for the four sample sizes considered herein). We rescaled the DGP to have unit variance and set σ equal to $0.25, 0.50, 1.00$, and 2.00, so that the expected R^2 for the unknown true model was $1/(1+\sigma^2)$ and was thus $0.95, 0.80, 0.50$, and 0.20, respectively.

The simulations used nested regression models with variables $\{x_{ij}, j = 1, \ldots, M_n\}$. We considered six estimators: (1) Mallows model averaging defined over kernel smoothed varying coefficient candidates ('MMA'), (2) smoothed AIC model averaging ('SAIC'), (3) smoothed BIC model averaging ('SBIC'), (4) AIC model selection ('AIC'), (5) BIC model selection ('BIC'), and (6) Mallows' C_p model selection. All bandwidths were selected via least-squares cross validation. To evaluate the estimators, we computed

the risk (expected squared error). We did this by computing means (medians) across 1,000 simulation draws.

The SAIC and SBIC weights for the $j = 1, 2, \ldots, M$ models are given by

$$w_j = \exp(-AIC_j/2) / \sum_{j=1}^{M_n} \exp(-AIC_j/2),$$
$$w_j = \exp(-BIC_j/2) / \sum_{j=1}^{M_n} \exp(-BIC_j/2)$$

where AIC_j and BIC_j are given by $\log(\hat{\sigma}_j^2) + 2n^{-1}\,\mathrm{trace}(\mathbf{P}_{(j)})$ and $\log(\hat{\sigma}_j^2) + n^{-1}\,\mathrm{trace}(\mathbf{P}_{(j)})\log(n)$, respectively. The C_p criterion is given by $\hat{\sigma}_j^2(n + 2\,\mathrm{trace}(\mathbf{P}_{(j)}))$ where $\hat{\sigma}_j^2 = n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_{i,j}^2$ and where the $\hat{\epsilon}_{i,j}$ are the residuals from jth model.

Let $H = (\hat{\mu}_{(1)} - y, \ldots, \hat{\mu}_{(M_n)} - y)$ and let $b = \{\mathrm{trace}(P_{(1)}\hat{\Omega}_{(M_n)}), \ldots, \mathrm{trace}(P_{(M_n)}\hat{\Omega}_{(M_n)})\}^T$, where $\hat{\Omega}_{(M_n)}$ is a diagonal matrix formed from the squared residuals from the model indexed by the largest j (i.e. M_n). We can rewrite $\hat{C}_n(w)$ as $\hat{C}_n(w) = w^T H^T H w + 2w^T b$, which is a quadratic function of the weight vector w and the optimization can be done by standard software packages such as the `R` package `quadprog` (code underlying this simulation can be found in Supplementary Material 2). Using the largest model to estimate the error covariance matrix is advocated by Hansen (2007) and Liu and

Okui (2013), and in small samples this approach performs admirably.

Simulation results are summarized in Table 1, which reports the mean relative MSE row normalized so that the method with lowest mean MSE has entry 1.00. R^2 is higher for smaller values of σ; for larger values of α the $\theta_j(z)$ coefficients decay more rapidly with j. MMA, SAIC, and SBIC are model averaging methods; AIC, BIC and C_p are model selection methods.

3.1 Discussion

Clearly no one method dominates over the range of sample sizes, signal to noise ratio, and range of parameter decay considered above. AIC and C_p have similar risk. If one considers the range of risk relative to the best performing method in any experiment (row of Table 1), it would appear that the proposed approach dominates its peers while, as n increases, it clearly emerges as the preferred approach. On the basis of these simulations, the proposed method ought to appeal to practitioners interested in model average estimators defined over the flexible and popular varying coefficient specification.

Table 1: Monte Carlo Simulation Mean Relative MSE (row normalized so that the method with lowest mean MSE has entry 1.00). R^2 is higher for smaller values of σ; for larger values of α the $\theta_j(z)$ coefficients decay more rapidly with j. MMA, SAIC, and SBIC are model averaging methods; AIC, BIC and C_p are model selection methods.

n	α	σ	MMA	SAIC	SBIC	AIC	BIC	C_p
50	0.10	0.25	1.01	1.31	1.36	1.00	1.59	1.02
50	0.10	0.50	1.00	1.15	1.18	1.05	1.50	1.06
50	0.10	1.00	1.09	1.00	1.00	1.26	1.37	1.26
50	0.10	2.00	1.41	1.03	1.00	1.77	1.21	1.75
50	0.25	0.25	1.00	1.36	1.43	1.02	1.47	1.03
50	0.25	0.50	1.00	1.10	1.13	1.07	1.42	1.08
50	0.25	1.00	1.20	1.00	1.00	1.41	1.43	1.40
50	0.25	2.00	1.51	1.04	1.00	1.93	1.24	1.90
50	0.50	0.25	1.00	1.22	1.28	1.07	1.30	1.08
50	0.50	0.50	1.09	1.00	1.01	1.23	1.34	1.22
50	0.50	1.00	1.39	1.02	1.00	1.68	1.47	1.66
50	0.50	2.00	1.63	1.05	1.00	2.12	1.24	2.09
100	0.10	0.25	1.00	1.26	1.29	1.00	1.61	1.01
100	0.10	0.50	1.00	1.15	1.18	1.03	1.53	1.04
100	0.10	1.00	1.02	1.00	1.01	1.14	1.38	1.14
100	0.10	2.00	1.24	1.01	1.00	1.57	1.19	1.56
100	0.25	0.25	1.00	1.33	1.39	1.02	1.54	1.03
100	0.25	0.50	1.00	1.13	1.16	1.06	1.48	1.06
100	0.25	1.00	1.09	1.00	1.00	1.26	1.45	1.26
100	0.25	2.00	1.33	1.02	1.00	1.75	1.24	1.73
100	0.50	0.25	1.00	1.22	1.30	1.07	1.46	1.08
100	0.50	0.50	1.05	1.00	1.02	1.19	1.38	1.19
100	0.50	1.00	1.26	1.01	1.00	1.54	1.43	1.53
100	0.50	2.00	1.41	1.03	1.00	1.91	1.16	1.90
200	0.10	0.25	1.00	1.22	1.25	1.00	1.45	1.00
200	0.10	0.50	1.00	1.15	1.17	1.02	1.46	1.02
200	0.10	1.00	1.00	1.02	1.03	1.07	1.41	1.07
200	0.10	2.00	1.10	1.00	1.00	1.33	1.29	1.32
200	0.25	0.25	1.00	1.30	1.35	1.01	1.47	1.01
200	0.25	0.50	1.00	1.15	1.18	1.04	1.48	1.04
200	0.25	1.00	1.03	1.00	1.01	1.14	1.45	1.13
200	0.25	2.00	1.17	1.01	1.00	1.46	1.38	1.46
200	0.50	0.25	1.00	1.23	1.30	1.06	1.56	1.06
200	0.50	0.50	1.01	1.00	1.02	1.14	1.44	1.14
200	0.50	1.00	1.15	1.00	1.00	1.38	1.46	1.38
200	0.50	2.00	1.21	1.02	1.00	1.59	1.30	1.59
400	0.10	0.25	1.00	1.21	1.23	1.00	1.32	1.00
400	0.10	0.50	1.00	1.16	1.17	1.00	1.37	1.00
400	0.10	1.00	1.00	1.06	1.06	1.04	1.41	1.03
400	0.10	2.00	1.06	1.00	1.00	1.20	1.39	1.20
400	0.25	0.25	1.00	1.30	1.34	1.00	1.35	1.00
400	0.25	0.50	1.00	1.18	1.20	1.02	1.44	1.02
400	0.25	1.00	1.00	1.02	1.03	1.08	1.46	1.08
400	0.25	2.00	1.10	1.00	1.00	1.31	1.51	1.31
400	0.50	0.25	1.00	1.27	1.34	1.04	1.57	1.04
400	0.50	0.50	1.00	1.04	1.07	1.11	1.53	1.11
400	0.50	1.00	1.10	1.00	1.00	1.31	1.56	1.31
400	0.50	2.00	1.14	1.01	1.00	1.45	1.46	1.44
Mean (all n)			1.10	1.10	1.11	1.25	1.42	1.25
Mean ($n = 50$)			1.19	1.11	1.12	1.38	1.38	1.38
Mean ($n = 100$)			1.12	1.10	1.11	1.29	1.40	1.29
Mean ($n = 200$)			1.06	1.09	1.11	1.19	1.43	1.19
Mean ($n = 400$)			1.03	1.10	1.12	1.13	1.45	1.13

4. Empirical Illustration

In what follows we report an estimate of a Mincer (earnings) equation using Wooldridge's (2002) 'wage1' data which contains $n = 526$ observations on a range of variables. We considered modeling expected (log) hourly wages ('*lwage*') based on a number of commonly employed predictors, namely

1. *educ*: years of education
2. *exper*: years potential experience
3. *tenure*: years with current employer
4. *female*: "Female" if female, "Male" otherwise
5. *married*: "Married" if Married, "Nonmarried" otherwise

We treated the predictors *educ*, *exper*, and *tenure* as belonging to X and *female* and *married* as belonging to Z. We considered varying coefficient models that differed in terms of the contents of X. Let d be the order of a (orthogonal) polynomial formed from each of *educ*, *exper*, and *tenure*. When $d = 1$ there are 3 columns in X (*educ*, *exper*, and *tenure*) and if we consider all possible combinations of the predictors taken 1, 2, and 3 at a time then there are $M = \binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 7$ candidate models. When $d = 2$ there are 6 columns in X hence $M = 63$ candidate models, and when $d = 3$ there are 9 columns in X hence $M = 511$ candidate models. We also considered standard nonparametric local constant ('LC'), nonparametric local linear

('LL'), and semiparametric varying coefficient ('VC') models defined over the full set of predictors by way of comparison; see Li and Racine (2007, Pages 60, 79, and 301, respectively) for details.

We conducted a simulation in which the data was repeatedly shuffled and split into two parts 1,000 times, based on an estimation sample of size $n_1 = 500$ and an independent validation sample of size $n_2 = 26$. For each estimation sample we fit the cross-validated semiparametric varying coefficient model and each of the parametric and nonparametric models listed above. All bandwidths were selected via least-squares cross validation. For each model we then computed predicted square error ('PSE') for the independent validation data set given by $\text{PSE} = n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{Y}_i)^2$ where $\hat{Y}_i$ is the prediction for a given model. The mean relative hold-out PSE is presented in Table 2, row normalized so that the method with lowest mean PSE has entry 1.00, while the mean PSE is presented in Table 3.

Table 2: Empirical Illustration Mean Relative PSE (row normalized so that the method with lowest mean PSE has entry 1.00). MMA, SAIC, and SBIC are model averaging methods; AIC, BIC and C_p are model selection methods; LC, LL, and VC are nonparametric and semiparametric models.

		Model Average			Model Selection			Model Specification		
d	M	MMA	SAIC	SBIC	AIC	BIC	C_p	LC	LL	VC
1	7	1.043	1.080	1.081	1.041	1.051	1.041	1.041	1.000	1.040
2	63	1.000	1.056	1.057	1.008	1.054	1.008	1.082	1.039	1.089
3	511	1.000	1.061	1.062	1.029	1.056	1.029	1.075	1.039	1.093

Table 3: Empirical Illustration Mean PSE. MMA, SAIC, and SBIC are model averaging methods; AIC, BIC and C_p are model selection methods; LC, LL, and VC are nonparametric and semiparametric models.

		Model Average			Model Selection			Model Specification		
d	M	MMA	SAIC	SBIC	AIC	BIC	C_p	LC	LL	VC
1	7	0.167	0.173	0.173	0.167	0.169	0.167	0.167	0.160	0.167
2	63	0.151	0.160	0.160	0.153	0.159	0.153	0.164	0.157	0.165
3	511	0.152	0.161	0.161	0.156	0.160	0.156	0.163	0.158	0.166

Table 2 reveals some interesting features. First, from row 1 (i.e., $d = 1$), when we average across models in which the parametric component X is linear, the fully nonparametric local linear estimator is the best performer, dominating both model averaging and model selection, which for some might be unexpected. However, when we move to a larger number of candidate models allowing for quadratic ($d = 2$) and cubic ($d = 3$) terms to enter in the parametric component X, this appears to be sufficient for the model averaging estimator to dominate its peers. Furthermore, Table 3 reveals that there is no further MSE improvement in either the selection or averaging methods when we move from $d = 2$ to $d = 3$, hence a relatively modest number of candidate models appears to be sufficient for the proposed model averaging method to dominate its peers.

5. Concluding Remarks

In this paper we present a semiparametric approach to model averaging that possesses a number of desirable features. Theoretical underpinnings are provided, and its finite-sample performance indicates that it ought to be of interest to practitioners who wish to tackle model uncertainty. An illustrative application indicates that the method is capable of delivering models with impressive approximation capabilities. In particular, it can be seen how averaging over a set of semiparametric models can outperform fully nonparametric specifications in applied settings. R code for implementing the proposed approach is presented in the Supplementary Material, and is available upon request from the authors.

Supplementary Materials

Proofs of the main theorems are provided in Supplementary Material 1, while R code can be found in Supplementary Material 2.

Acknowledgements

We are indebted to Peter Hall for his deep and broad contributions to the statistics community. His contributions to semi- and nonparametric estimation have had a profound impact on the field, and we dedicate this paper to his memory. Racine would like to gratefully acknowledge

support from the Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca). Qi Li's research is partially supported by China National Science Foundation, projects #71601130 and #71133001.

References

Akaike, H. (1970). Statistical Predictor Identification. *Annals of the Institute of Statistics and Mathematics.* 22, 203-217.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Second International Symposium on Information Theory.* (Petroc, B. and F. Csake eds).Akademiai Kiado, Budapest, 267-281.

Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association.* 109(505), 254-265.

Andrews, D. W. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics.* 47(2), 359-377.

Beran, R. and P. Hall (1992). Estimating Coefficient Distributions in Random Coefficient Regressions. *The Annals of Statistics.* 20(4), 1970-1984.

Buckland, S. T., K. P. Burnhamn and N. H. Augustin (1997). Model Selection: An Integral Part of Inference. *Biometrics.* 53, 603618.

Cai, Z., J. Fan and Q. W. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association.* 95, 941-956.

Claeskens, G. and N. L. Hjort (2003). The Focused Information Criterion. *Journal of the American Statistical Association.* 98, 900-916.

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*. 13, 377-403.

Hall, P. G. and J. S. Racine (2015). Infinite order cross-validated local polynomial regression. *Journal of Econometrics*. 185(2), 510-525.

Hall, P., Q. Li and J. S. Racine (2007). Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors. *The Review of Economics and Statistics*. 89, 784-789.

Hall, P., J. S. Racine and Q. Li (2004). Cross-Validation and the Estimation of Conditional Probability Densities. *Journal of the American Statistical Association*. 99(468), 1015-1026.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*. 75(4), 1175-1189.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*. 5(3), 495-530.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics*. 167(1), 38-46.

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

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*. 14, 382417.

Li, Q., C. J. Huang, D. Li and T. T. Fu (2002). Semiparametric Smooth Coefficient Models. *Journal of Business and Economics Statistics*. 20, 412-422.

Li, Q. and J. S. Racine. (2007). Nonparametric Econometrics: Theory and Practice. *Princeton University Press*.

Liu, Q. and R. Okui (2013). Heteroscedasticity-robust C_p model averaging. *The Econometrics Journal*. 16(3), 463-472.

Liu, Q, R. Okui and A. Yoshimura (2016). Generalized least squares model averaging. *Econometric Reviews*. 1-16.

Lu, X. and L. Su (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics*. 188(1), 40-58.

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

Robinson, P. M. (1988). Root-N consistent semiparametric regression. *Econometrica*. 56, 931-954.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*. 6, 461-464.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*. 413–436.

Stone, C. J. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion). *Journal of the Royal Statistical Society*. 36, 111-147.

Wan, A. T., X. Zhang and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*. 156(2), 277-283.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent

variables. *Theory of Probability & Its Applications*. 5(3), 302–305.

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. *MIT Press*. Cambridge.

Zhang, Xinyu and Wendun Wang (2015). *Optimal Model Averaging Estimation for Partially Linear Models*, Rechnical report. URL: https://ssrn.com/abstract=2948380.

Zhang, X., D. Yu, G. Zou and H. Liang. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of American Statistical Association.* ,In Press.

Zhang, X., G. Zou and R.J. Carroll. (2015). Model averaging based on Kullback-Leibler distance. *Statistic Sinica.* ,Vol.25 pp. 1583–1598.

Zhao, S, X. Zhang and Y. Gao. (2016). Model averaging with averaging covariance matrix. *Economics Letters.* ,Vol.145 pp. 214–217.

Cong Li, School of Economics, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai, 200433, PR China.

E-mail: (li.cong@mail.shufe.edu.cn)

Qi Li, ISEM, Capital University of Economics and Business, Beijing, 100070, PR China; Department of Economics, Texas A&M University, College Station, TX 77843, USA.

E-mail: (qi-li@tamu.edu)

Jeffrey S. Racine, Department of Economics and Graduate Program in Statistics, McMaster University; Department of Economics and Finance, La Trobe University; Info-Metrics Institute, American University; Rimini Center for Economic Analysis; Center for Research in Econometric Analysis of Time Series (CREATES), Aarhus University.

E-mail: (racinej@mcmaster.ca)

Daiqiang Zhang, Department of Economics, University at Albany, SUNY, Albany, NY 12222, USA.

E-mail: (dzhang6@albany.edu)