Journal of Econometrics 192 (2016) 139–151

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Model averaging based on leave-subject-out cross-validation

Yan Gao [a,b], Xinyu Zhang [b,c,*], Shouyang Wang [b], Guohua Zou [b,d]

[a] Department of Statistics, College of Science, Minzu University of China, Beijing 100081, China
[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[c] ISEM, Capital University of Economics and Business, Beijing 100070, China
[d] School of Mathematical Science, Capital Normal University, Beijing 100037, China

ARTICLE INFO

Article history:
Received 23 August 2014
Received in revised form
16 April 2015
Accepted 20 July 2015
Available online 28 December 2015

JEL classification:
C51
C52

Keywords:
Asymptotic optimality
Leave-subject-out cross-validation
Longitudinal data
Model averaging
Time series

ABSTRACT

This paper develops a frequentist model averaging method based on the leave-subject-out cross-validation. This method is applicable not only to averaging longitudinal data models, but also to averaging time series models which can have heteroscedastic errors. The resulting model averaging estimators are proved to be asymptotically optimal in the sense of achieving the lowest possible squared errors. Both simulation study and empirical example show the superiority of the proposed estimators over their competitors.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Model averaging (MA), a smoothed extension of model selection (MS), generally yields a lower risk than model selection. There are many studies on Bayesian model averaging. Hoeting et al. (1999) provided a comprehensive review in this direction. Recent years have also witnessed a booming development of frequentist model averaging methods such as weighting strategy based on the scores of information criteria (Buckland et al., 1997; Hjort and Claeskens, 2003, 2006; Zhang and Liang, 2011; Zhang et al., 2012; Xu et al., 2014), the adaptive regression by mixing by Yang (2001), Mallows model averaging (MMA) by Hansen (2007), and optimal mean squared error averaging by Liang et al. (2011). Recently, taking heteroscedasticity into consideration, Hansen and Racine (2012) proposed a jackknife model averaging (JMA) method that selects weights by minimizing a leave-one-out cross-validation criterion. The JMA estimator performs quite well in cross-sectional data. However, for longitudinal data, there generally exists within-subject correlation in error terms, thus the JMA method may not be appropriate. In the current paper, we develop a model averaging estimator called leave-subject-out model averaging (LsoMA) estimator for longitudinal data models.

There exists a rich literature on longitudinal data models. For an overview of parametric longitudinal data models, one can refer to Arellano (2003), Hsiao (2003) and Baltagi (2005). Nonparametric (Rice and Silverman, 1991; Fan and Zhang, 2000; Welsh et al., 2002; Zhu et al., 2008) and semiparametric longitudinal data models (Zeger and Diggle, 1994; Zhang et al., 1998; Lin and Ying, 2001) have also been considered. Penalized model selection methods are commonly used in nonparametric and semiparametric models. In the current paper, we use a quadratic penalty based on smoothing splines. The popular nonquadratic penalties, such as the least absolute shrinkage and selection operator (Tibshirani, 1996), hard thresholding (Antoniadis, 1997; Fan, 1997), and the smoothly clipped absolute deviation penalties (Fan and Li, 2001) can also be utilized here. For all these methods, tuning parameters need to be selected. Rice and Silverman (1991) introduced the leave-subject-out cross-validation (LsoCV) to select tuning parameters. This method has been widely used in longitudinal data model since then. For example, Xu and Huang (2012) utilized the LsoCV to select variables in the semiparametric longitudinal data model, and proved the asymptotic optimality of their approach. Further, they developed an efficient computation procedure for the LsoCV.

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: xinyu@amss.ac.cn (X. Zhang).

http://dx.doi.org/10.1016/j.jeconom.2015.07.006
0304-4076/© 2015 Elsevier B.V. All rights reserved.

The current paper focuses on model averaging. By using our model averaging method, the estimators based on different covariates or different tuning parameters are asymptotically optimally combined, i.e., under some regularity conditions, our model averaging estimator minimizes predictive squared error in large sample cases. A related work dealing with longitudinal data was done by Zhang et al. (2014). They developed a model averaging method to combine forecasts from linear mixed-effects models. In the current paper, to be more general, we allow each candidate model to contain non-parametric component.

On the other hand, series dependence always exists in time series models. A natural idea is to treat those with high correlation as a subject, and thus we can use the LsoMA method to combine time series models. In this paper, we will show that the resulting LsoMA estimator is also asymptotically optimal. In Cheng and Hansen (2015), the LsoCV criterion was suggested to average forecasts for factor-augmented regressions, but the corresponding asymptotic optimality was not achieved in that work.

We do a Monte Carlo study to compare the finite sample performance of the proposed LsoMA method with others including model selection methods by AIC, BIC and LsoCV, and model averaging methods by smoothed AIC and smoothed BIC (Buckland et al., 1997), and JMA in both longitudinal data model and time series model. Simulation results indicate that the LsoMA estimator performs better than its competitors in most cases. We also conduct an empirical study on the Chinese consumer price index, which shows that our method has better forecasting performance than the commonly used model selection and averaging methods.

The remainder of this paper is organized as follows. Section 2 proposes the LsoMA estimator for longitudinal data model and develops its asymptotic optimality theory. Section 3 studies the LsoMA method for time series models. Section 4 numerically compares our LsoMA estimators with those obtained from some commonly used model selection and model averaging methods. Section 5 conducts an empirical study. Section 6 concludes. The proofs are relegated to Appendix.

2. Leave-subject-out model averaging for longitudinal data models

2.1. Model framework

Suppose that $(y_{ij}, \mathbf{x}_{ij}), j = 1, \ldots, T_i$, are observations for subject i, $i = 1, \ldots, n$. Let $\mathbf{Y}_i = (y_{i1}, \ldots, y_{iT_i})'$, $\widetilde{\mathbf{X}}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i})'$ and $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1', \ldots, \widetilde{\mathbf{X}}_n')'$. We consider the following semiparametric model for longitudinal data

$$\mathbf{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$
$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT_i})',$$
$$\mu_{ij} = E(y_{ij}|\widetilde{\mathbf{X}}_i) = \mathbf{x}_{ij,0}'\boldsymbol{\beta}_0 + \sum_{l=1}^{L} f_l(x_{ij,l}), \quad j = 1, \ldots, T_i,$$

where $\mathbf{x}_{ij,0}$ contains variables of parametric component, $x_{ij,1}, \ldots, x_{ij,L}$ are variables of nonparametric component, $\boldsymbol{\beta}_0$ is the coefficient vector of the linear component, $f_1, \ldots, f_L$ are smooth functions, and $\boldsymbol{\varepsilon}_i$'s are independent disturbances with $E(\boldsymbol{\varepsilon}_i|\widetilde{\mathbf{X}}_i) = 0$ and $E(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\widetilde{\mathbf{X}}_i) = \boldsymbol{\Sigma}_i$. We can use a basis expansion to approximate each f_l. Then, there exist a design matrix $\mathbf{X}_i$ and an unknown parameter vector $\boldsymbol{\beta}$ such that $\boldsymbol{\mu}_i \approx \mathbf{X}_i\boldsymbol{\beta}$. Specifically, $\mathbf{X}_i = (\mathbf{X}_{1,i}^*, \mathbf{X}_{2,i}^*)$ consists of two parts: the linear regression variables matrix $\mathbf{X}_{1,i}^*$, and the basis matrix $\mathbf{X}_{2,i}^*$ used to approximate the nonparametric component.

We estimate $\boldsymbol{\beta}$ by minimizing the following penalized weighted least squares (Xu and Huang, 2012)

$$\mathrm{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{l=1}^{L}\lambda_l\boldsymbol{\beta}'\mathbf{F}_l\boldsymbol{\beta},$$

where $\mathbf{V}_i$'s are working covariance matrices, $\mathbf{F}_l$ is a positive semi-definite matrix such that $\boldsymbol{\beta}'\mathbf{F}_l\boldsymbol{\beta}$ serves as a roughness penalty for f_l, and $\lambda_1, \ldots, \lambda_L$ are tuning parameters. In the current paper, penalties are put only on the nonlinear parts, so $\mathbf{F}_l$ is a block diagonal matrix with the block corresponding to the linear part being $\mathbf{0}$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)'$, $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n')'$, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_n')'$, and $\mathbf{V} = \mathrm{diag}\{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$. The estimator of $\boldsymbol{\beta}$ can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \sum_{l=1}^{L}\lambda_l\mathbf{F}_l\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

There are many methods for constructing the basis matrix $\mathbf{X}_2^* \equiv (\mathbf{X}_{2,1}^{*\prime}, \ldots, \mathbf{X}_{2,n}^{*\prime})'$ and the penalty matrix $\mathbf{F}_l$. For example, one can use the spline basis to generate $\mathbf{X}_2^*$. Linear spline is the simplest method but has a sharp corner disadvantage. Quadratic spline basis can remedy this disadvantage as it has a continuous first derivative. The truncated power basis of degree higher than two provides more complex regression functions. However, it may lead to numerical instability due to its nonorthogonality, which can be overcome by the B-spline basis. For the penalty matrix $\mathbf{F}_l$, we can utilize the squared second-difference penalty, the squared second derivative penalty or the thin-plate splines penalty. More details on the basis and penalty matrices can be found in Green and Silverman (1994) and Ruppert et al. (2003).

Following Claeskens et al. (2009), we take penalized B-spline as an example to show how to construct the basis and penalty matrices. For simplicity, we consider the model with only one nonparametric covariate, so that there is only one tuning parameter λ and $L = 1$. Let r be the order of B-splines. Define a sequence of knots on the interval $[I_{low}, I_{up}]$: $I_{low} = m_{-(r-1)} = \cdots = m_0 < m_1 < \cdots < m_{K_n} < m_{K_n+1} = \cdots = m_{K_n+r} = I_{up}$. Basis functions can be expressed as

$$B_{j,1}(x) = \begin{cases} 1, & m_j \le x < m_{j+1}, \\ 0, & \text{otherwise}, \end{cases}$$

$$B_{j,r}(x) = \frac{x - m_j}{m_{j+r-1} - m_j}B_{j,r-1}(x) + \frac{m_{j+r} - x}{m_{j+r} - m_{j+1}}B_{j+1,r-1}(x),$$

for $j = -(r-1), \ldots, K_n$. Then, the kth row of $\mathbf{X}_{2,i}^*$ is $(B_{-(r-1),r}(x_{ik,1}), \ldots, B_{K_n,r}(x_{ik,1}))$. The penalty term can be written as $\lambda\boldsymbol{\beta}_2'\boldsymbol{\Delta}_q'\mathbf{R}\boldsymbol{\Delta}_q\boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_2$ is the coefficient vector of $\mathbf{X}_2^*$, $\mathbf{R}$ is a $(K_n + r - q) \times (K_n + r - q)$ matrix with its ij element $R_{ij} = \int_{I_{low}}^{I_{up}} B_{j,r-q}(x)B_{i,r-q}(x)dx$, and $\boldsymbol{\Delta}_q$ is a matrix of qth order difference operator. If the knots are equidistant, i.e., $m_j - m_{j-1} = \delta$ for $j = 1, \ldots, K_n + 1$, then $\boldsymbol{\Delta}_q$ can be expressed in terms of the qth backward difference operator ∇_q, i.e., $\boldsymbol{\Delta}_q = \delta^{-q}\nabla_q$. Each element of the matrix ∇_q is defined recursively via $\nabla_q = \nabla_1(\nabla_{q-1})$ and $\nabla_1\beta_j = \beta_j - \beta_{j-1}$. If we take $q = r - 1$, then $\mathbf{R}$ reduces to a diagonal matrix with the diagonal element δ.

The working covariance matrices $\mathbf{V}_i$'s are generally estimated based on the working correlation structures of $\boldsymbol{\varepsilon}_i$'s. In practice, compound symmetry and autoregressive structures are commonly used working correlation structures. As commented by Liang and Zeger (1986), a possibly misspecified correlation structure also has a potential to improve the estimation efficiency over a method that completely ignores the within-subject correlation. Diggle et al. (2002) provided details on the choice of working correlation structure for longitudinal data. In the current paper, following Xu and Huang (2012), we set $\mathbf{V}_i$'s to be the identity matrices at first, based on which the model is estimated to get residuals, and then $\mathbf{V}_i$'s are estimated using these residuals.

2.2. Model averaging criterion

Assume that candidate estimators differ from each other in regressors and/or tuning parameters. Let $\mathcal{X} = \{\mathbf{X}^{(1)}, \ldots,$

$\mathbf{X}^{(M)}\}$[1] and $\mathcal{K} = \{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(K)}\}$ be the candidate sets of regressor matrices and tuning parameters for f_l's, respectively. Thus, the total number of candidate models is $S = MK$. For the sth candidate model, denote $\mathbf{X}^{(s)} = (\mathbf{X}_1^{(s)\prime}, \ldots, \mathbf{X}_n^{(s)\prime})'$ as the regressor matrix which is assumed to be of full column rank and $\boldsymbol{\lambda}^{(s)} = (\lambda_1^{(s)}, \ldots, \lambda_L^{(s)})'$ as the tuning parameter vector. Each pair of $\{\mathbf{X}^{(s)}, \boldsymbol{\lambda}^{(s)}\}$ is chosen from $\mathcal{X} \times \mathcal{K}$. So

$$\widehat{\boldsymbol{\mu}}^{(s)} = \mathbf{X}^{(s)}\left(\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1}\mathbf{X}^{(s)} + \sum_{l=1}^{L}\lambda_l^{(s)}\mathbf{F}_l^{(s)}\right)^{-1}\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1}\mathbf{Y} \equiv \mathbf{P}^{(s)}\mathbf{Y},$$

where $\mathbf{V}_{(s)} = \mathrm{diag}\{\mathbf{V}_{(s)1}, \ldots, \mathbf{V}_{(s)n}\}$ is the working covariance matrix, and $\mathbf{F}_l^{(s)}$ is the roughness penalty matrix under the sth candidate model.

Let $\mathbf{w} = (w_1, \ldots, w_S)'$ be a weight vector in the unit simplex in R^S

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

Then the model averaging estimator of $\boldsymbol{\mu}$ can be expressed as

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^{S} w_s\widehat{\boldsymbol{\mu}}^{(s)} = \sum_{s=1}^{S} w_s\mathbf{P}^{(s)}\mathbf{Y} \equiv \mathbf{P}(\mathbf{w})\mathbf{Y},$$

where $\mathbf{P}(\mathbf{w}) = \sum_{s=1}^{S} w_s\mathbf{P}^{(s)}$. Let $\boldsymbol{\Sigma} = \mathrm{diag}\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n\}$. The squared error is $L_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$ and the risk is

$$R_n(\mathbf{w}) = E(L_n(\mathbf{w})|\widetilde{\mathbf{X}}) = \|\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2 + tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma}), \quad (1)$$

with $\mathbf{A}(\mathbf{w}) = \mathbf{I} - \mathbf{P}(\mathbf{w})$.

As the observations from the same subject have within-subject correlation, we use a leave-subject-out method to select weights. Let

$$\mathbf{Y}_{[-i]} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_{i-1}', \mathbf{Y}_{i+1}', \ldots, \mathbf{Y}_n')',$$
$$\mathbf{X}_{[-i]}^{(s)} = (\mathbf{X}_1^{(s)\prime}, \ldots, \mathbf{X}_{i-1}^{(s)\prime}, \mathbf{X}_{i+1}^{(s)\prime}, \ldots, \mathbf{X}_n^{(s)\prime})',$$

and $\mathbf{V}_{(s)[-i]}$ be a matrix with the ith diagonal block of $\mathbf{V}_{(s)}$ deleted. The leave-subject-out estimator of $\boldsymbol{\mu}_i$ is

$$\widetilde{\boldsymbol{\mu}}_i^{(s)} = \mathbf{X}_i^{(s)}\left(\mathbf{X}_{[-i]}^{(s)\prime}\mathbf{V}_{(s)[-i]}^{-1}\mathbf{X}_{[-i]}^{(s)} + \sum_{l=1}^{L}\lambda_l^{(s)}\mathbf{F}_l^{(s)}\right)^{-1}\mathbf{X}_{[-i]}^{(s)\prime}\mathbf{V}_{(s)[-i]}^{-1}\mathbf{Y}_{[-i]}.$$

Denote $\widetilde{\boldsymbol{\mu}}^{(s)} = (\widetilde{\boldsymbol{\mu}}_1^{(s)\prime}, \ldots, \widetilde{\boldsymbol{\mu}}_n^{(s)\prime})'$ and $\widetilde{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^{S} w_s\widetilde{\boldsymbol{\mu}}^{(s)}$. The LsoCV weight choice criterion is given by

$$\mathrm{LsoCV}(\mathbf{w}) = \|\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})\|^2. \quad (2)$$

By minimizing (2), we can obtain $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}}\mathrm{LsoCV}(\mathbf{w})$. The estimator $\widehat{\boldsymbol{\mu}}(\widehat{\mathbf{w}})$ is referred to as the leave-subject-out model averaging (LsoMA) estimator of $\boldsymbol{\mu}$ hereafter.

The definition of $\widetilde{\boldsymbol{\mu}}^{(s)}$ indicates that we need to calculate $\widetilde{\boldsymbol{\mu}}_1^{(s)}, \ldots, \widetilde{\boldsymbol{\mu}}_n^{(s)}$ separately for each candidate model. This is cumbersome, and so it is necessary to develop a computational shortcut form to calculate them. Denote $\widehat{\mathbf{Y}}_i^{(s)} = \mathbf{L}_i\mathbf{P}^{(s)}\mathbf{Y}$, where $\mathbf{L}_i = (\mathbf{0}_{T_i\times T_1}, \ldots, \mathbf{0}_{T_i\times T_{i-1}}, \mathbf{I}_{T_i}, \mathbf{0}_{T_i\times T_{i+1}}, \ldots, \mathbf{0}_{T_i\times T_n})$ and $\mathbf{I}_{T_i}$ is the $T_i \times T_i$ identity matrix.[2] Let $\mathbf{P}_{ii}^{(s)}$ be the ith diagonal block matrix of $\mathbf{P}^{(s)}$. From Xu and Huang (2012), in the sth candidate model,

$$\mathbf{Y}_i - \widetilde{\boldsymbol{\mu}}_i^{(s)} = (\mathbf{I}_{T_i} - \mathbf{P}_{ii}^{(s)})^{-1}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i^{(s)}) \quad (3)$$

and

$$(\mathbf{I}_{T_i} - \mathbf{P}_{ii}^{(s)})^{-1} = \mathbf{I}_{T_i} + \mathbf{P}_{ii}^{(s)} + (\mathbf{P}_{ii}^{(s)})^2 + (\mathbf{P}_{ii}^{(s)})^3 + \cdots \equiv \mathbf{I}_{T_i} + \mathbf{D}_i^{(s)}.$$

Let $\mathbf{D}^{(s)} = \mathrm{diag}(\mathbf{D}_1^{(s)}, \ldots, \mathbf{D}_n^{(s)})$, $\mathbf{A}^{(s)} = \mathbf{I} - \mathbf{P}^{(s)}$, $\mathbf{Q}^{(s)} = \mathbf{D}^{(s)}\mathbf{A}^{(s)}$, and $\mathbf{Q}(\mathbf{w}) = \sum_{s=1}^{S} w_s\mathbf{Q}^{(s)}$. Thus, we have

$$\mathrm{LsoCV}(\mathbf{w}) = \mathbf{Y}'(\mathbf{A}(\mathbf{w}) + \mathbf{Q}(\mathbf{w}))'(\mathbf{A}(\mathbf{w}) + \mathbf{Q}(\mathbf{w}))\mathbf{Y}. \quad (4)$$

Hansen and Racine (2012) and Zhang et al. (2013) studied the JMA method in linear cross-sectional case. If each subject has only one observation and the least squares estimator is used, i.e., $T_i = 1$ $(i = 1, \ldots, n)$, $\mathbf{V}_{(s)} = \mathbf{I}$ and $\lambda_l = 0$, then the criterion LsoCV($\mathbf{w}$) will reduce to the jackknife criterion in their works.

2.3. Asymptotic optimality

To show the asymptotic optimality of the proposed estimator $\widehat{\boldsymbol{\mu}}(\widehat{\mathbf{w}})$, we will list some regularity conditions first. Let $N = \sum_{i=1}^{n} T_i$, $k_s = \mathrm{rank}(\mathbf{X}^{(s)})$, $S^* = \arg\max_{s\in\{1,\ldots,S\}} k_s$, $\xi_n = \inf_{\mathbf{w}\in\mathcal{H}} R_n(\mathbf{w})$, $\mathbf{e}_i = \boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\varepsilon}_i$, and $\mathbf{u}_{ij}$ be a $T_i \times 1$ vector such that $\mathbf{u}_{ij}'\mathbf{u}_{ij} = 1$ and $\mathbf{u}_{ij}'\mathbf{u}_{ik} = 0$ if $j \neq k$, where $j, k = 1, \ldots, T_i$ and $i = 1, \ldots, n$. Denote $\bar{\sigma}(\cdot)$ and $c(\cdot)$ as the largest singular value and condition number of a matrix, respectively. Note that $\mathbf{P}^{(s)}$ is not symmetric unless $\mathbf{V}_{(s)} = \mathbf{I}$. We define a symmetric matrix $\widetilde{\mathbf{P}}^{(s)} = \mathbf{V}_{(s)}^{-1/2}\mathbf{P}^{(s)}\mathbf{V}_{(s)}^{1/2}$. Let $\widetilde{\mathbf{P}}_{ii}^{(s)}$ and $(\widetilde{\mathbf{P}}^{(s)})_{ii}^2$ be the ith diagonal block of $\widetilde{\mathbf{P}}^{(s)}$ and $(\widetilde{\mathbf{P}}^{(s)})^2$, respectively. All limiting processes discussed in this section are with respect to $n \to \infty$.

Condition 1. For some constant $C_1 < \infty$, $E[(\mathbf{u}_{ij}'\mathbf{e}_i)^4|\widetilde{\mathbf{X}}_i] \le C_1$ holds for all $j = 1, \ldots, T_i$ and $i = 1, \ldots, n$ *a.s.*

Condition 2. (**i**) $\max_{1\le i\le n}\max_{1\le s\le S} tr\mathbf{P}_{ii}^{(s)} = O(k_{S^*}/n)$ and $k_{S^*} = o(n)$ *a.s.*;
(**ii**) $\max_{1\le i\le n}\max_{1\le s\le S} tr(\widetilde{\mathbf{P}}^{(s)})_{ii}^2 = o(1)$ *a.s.*

Condition 3. $\|\boldsymbol{\mu}\|^2 = O(N)$ *a.s.*

Condition 4. $\xi_n^{-2}\sum_{s=1}^{S} R_n(\mathbf{w}_s^0) \xrightarrow{a.s.} 0$, where $\mathbf{w}_s^0$ is an $S \times 1$ vector in which the sth element is one and the others are zeros.

Condition 5. $\bar{\sigma}(\boldsymbol{\Sigma}) \le C_2$, where $C_2 < \infty$ is a constant.

Condition 6. $Nk_{S^*}S(n\xi_n)^{-1} \xrightarrow{a.s.} 0$.

Condition 7. For some constant $C_3 < \infty$, $\max_{1\le s\le S} c(\mathbf{V}_{(s)}) \le C_3$.

Conditions 1 and 2 are similar to Conditions 1 and 2 of Xu and Huang (2012). Condition 1 is a mild moment condition that requires $\mathbf{e}_i$ have uniformly bounded fourth moment. Condition 2 means that for each candidate model, there should not be any dominant or extremely influential subjects (Xu and Huang, 2012). In the linear cross-sectional case, Condition 2(i) reduces to condition (24) of Zhang et al. (2013) and Condition 2(ii) is equivalent to Condition 2(i) because $\widetilde{\mathbf{P}}^{(s)}$ is symmetric and idempotent in this case. Conditions 3–5 are the same as the conditions (21), (23) and (12) of Zhang et al. (2013), respectively.

[1] Usually, all the possible combinations of variables in linear components are considered. But if M is very large, then the computational burden is heavy. In this case, a model screening step before model averaging is necessary. An approach for screening models proposed by Claeskens et al. (2006) is to utilize the backward or forward regression to order variables and then use nested sets of variables. Another method proposed by Yuan and Yang (2005) is a 'top m' model screening with the aid of information criterion (AIC or BIC), i.e., the method excludes candidate models whose information criterion values are larger than the mth smallest one. From the perspective of computational burden, the first method appears more attractive than the second one, because it does not need to calculate the information criterion values for every model. However, this method takes only one model for each size, with the selected models nested, and this may lead to a loss of information.

[2] In some places, we do not indicate the dimension of the identity matrix, because it can be obtained from the dimensions of other matrices.

Condition 6 is developed from condition (22) of Zhang et al. (2013). Condition 7 requires that all working covariance matrices are not nearly singular and it is naturally satisfied in the linear cross-sectional case. Xu and Huang (2012) also suggested choosing a working covariance matrix which is not singular.

Theorem 2.1. *Under Conditions 1–7, we have*

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} L_n(\mathbf{w})} \xrightarrow{p} 1. \tag{5}$$

Proof. See Appendix.

3. Leave-subject-out model averaging for time series models

The time series models often have series dependence, so here we apply leave-subject-out cross-validation to averaging such models. In this section, the sample size N is replaced with T. Consider the following partially linear autoregressive model

$$y_t = \mu_t + \varepsilon_t = f(y_{t-1}) + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t, \quad t = 1, \ldots, T, \tag{6}$$

where the first component is possibly nonlinear, and heteroscedasticity is allowed among ε_t's. Assume that p is finite. Here, for simplicity, we consider the case with only one nonlinear component $f(y_{t-1})$. If we use another lag variable $y_{t-\widetilde{p}}$ instead of y_{t-1} in the nonlinear component, the following theoretical property keeps the same.[3] The theory for the case of more nonlinear components can be derived along the line of this simple case. We use the B-spline method mentioned in Section 2.1 to construct the design matrix, and thus the tth row of the design matrix is $\mathbf{X}_t = (B_{-(r-1),r}(y_{t-1}), \ldots, B_{K_T,r}(y_{t-1}), y_{t-2}, \ldots, y_{t-p})'$.

We consider a series of approximating models. These models differ in lag orders. As in Section 2.2, denote $\mathbf{X}_t^{(s)} = (B_{-(r-1),r}(y_{t-1}), \ldots, B_{K_T,r}(y_{t-1}), y_{t-2}, \ldots, y_{t-p_s})'$, and let $\mathbf{X}^{(s)} = (\mathbf{X}_1^{(s)}, \ldots, \mathbf{X}_T^{(s)})'$ and $\boldsymbol{\lambda}^{(s)}$ be the regressor matrix and tuning parameter for the sth candidate model, respectively. Then model s can be expressed as

$$\mathbf{Y} = \mathbf{X}^{(s)}\boldsymbol{\alpha}^{(s)} + \boldsymbol{\varepsilon}.$$

Setting the working covariance matrix to be $\mathbf{I}$, we obtain the following estimator of $\boldsymbol{\mu}$,

$$\widehat{\boldsymbol{\mu}}^{(s)} = \mathbf{X}^{(s)}\left(\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)} + \sum_{l=1}^{L}\lambda_l^{(s)}\mathbf{F}_l^{(s)}\right)^{-1}\mathbf{X}^{(s)\prime}\mathbf{Y} \equiv \mathbf{P}^{(s)*}\mathbf{Y}.$$

When calculating the leave-subject-out estimator of μ_t, we treat the lag variables with high correlation with y_t as a subject. There are a variety of ways to measure correlation such as the autocorrelation function, partial autocorrelation function (PACF), and mixing coefficient. Here we take the second measurement as an example to present how to determine a subject. We first calculate the PACF of the series. In general, it will decrease to 0 as the lag order increases. When the absolute value of PACF begins to fall below a given critical value c_r, we record the lag order a_0. Then a subject is formed based on a_0.

[3] In empirical analysis, one can choose the nonlinear variable according to the data features.

For the sth candidate model, let p_s be the autoregressive order, $b_s = \max\{a_0, p_s\}$,

$$\boldsymbol{\Phi}_t^{(s)} = \begin{cases} (\mathbf{I}_{b_s+t}, \mathbf{0}_{(b_s+t)\times(T-b_s-t)}), \\ \quad 1 \le t \le b_s+1, \\ (\mathbf{0}_{(2b_s+1)\times(t-b_s-1)}, \mathbf{I}_{2b_s+1}, \mathbf{0}_{(2b_s+1)\times(T-b_s-t)}), \\ \quad b_s+2 \le t \le T-b_s-1, \\ (\mathbf{0}_{(T+b_s-t+1)\times(t-b_s-1)}, \mathbf{I}_{T+b_s-t+1}), \\ \quad T-b_s \le t \le T, \end{cases}$$

$\mathbf{Y}_{[-t]}^{(s)} = \boldsymbol{\Phi}_t^{(s)}\mathbf{Y}$, $\mathbf{X}_{[-t]}^{(s)} = \boldsymbol{\Phi}_t^{(s)}\mathbf{X}^{(s)}$, and $\boldsymbol{\mu}_{[-t]}^{(s)} = \boldsymbol{\Phi}_t^{(s)}\boldsymbol{\mu}$. Under the sth candidate model, we delete $\{\mathbf{Y}_{[-t]}^{(s)}, \mathbf{X}_{[-t]}^{(s)}\}$ from $\{\mathbf{Y}, \mathbf{X}^{(s)}\}$ to get the leave-subject-out estimator $\widetilde{\widetilde{\mu}}_t^{(s)}$ of μ_t. Denote

$$\boldsymbol{\pi}_t^{(s)} = \begin{cases} (\underbrace{0, \ldots, 0}_{t-1}, 1, \underbrace{0, \ldots, 0}_{b_s}), & 1 \le t \le b_s+1, \\ (\underbrace{0, \ldots, 0}_{b_s}, 1, \underbrace{0, \ldots, 0}_{b_s}), & b_s+2 \le t \le T-b_s-1, \\ (\underbrace{0, \ldots, 0}_{b_s}, 1, \underbrace{0, \ldots, 0}_{T-t}), & T-b_s \le t \le T. \end{cases}$$

Then it is readily seen that $\mu_t = \boldsymbol{\pi}_t^{(s)}\boldsymbol{\mu}_{[-t]}^{(s)}$. Let $\widehat{\mathbf{Y}}_{[-t]}^{(s)} = \boldsymbol{\Phi}_t^{(s)}\mathbf{P}^{(s)*}\mathbf{Y}$, $\mathbf{P}_{tt}^{(s)*} = \boldsymbol{\Phi}_t^{(s)}\mathbf{P}^{(s)*}\boldsymbol{\Phi}_t^{(s)\prime}$, and $\widetilde{\boldsymbol{\mu}}_{[-t]}^{(s)} = \mathbf{Y}_{[-t]}^{(s)} - (\mathbf{I} - \mathbf{P}_{tt}^{(s)*})^{-1}(\mathbf{Y}_{[-t]}^{(s)} - \widehat{\mathbf{Y}}_{[-t]}^{(s)})$. From (3), it is clear that $\widetilde{\boldsymbol{\mu}}_{[-t]}^{(s)}$ is the leave-subject-out estimator of $\boldsymbol{\mu}_{[-t]}^{(s)}$. So $\widetilde{\widetilde{\mu}}_t^{(s)}$ is just an element of $\widetilde{\boldsymbol{\mu}}_{[-t]}^{(s)}$ satisfying

$$\widetilde{\widetilde{\mu}}_t^{(s)} = \boldsymbol{\pi}_t^{(s)}\widetilde{\boldsymbol{\mu}}_{[-t]}^{(s)} = \boldsymbol{\pi}_t^{(s)}\left[\mathbf{Y}_{[-t]}^{(s)} - (\mathbf{I} - \mathbf{P}_{tt}^{(s)*})^{-1}(\mathbf{Y}_{[-t]}^{(s)} - \widehat{\mathbf{Y}}_{[-t]}^{(s)})\right]. \tag{7}$$

Define $N^* = T - b_s^2 + (2T-1)b_s$, $\widetilde{\widetilde{\boldsymbol{\mu}}}^{(s)} = (\widetilde{\widetilde{\mu}}_1^{(s)}, \ldots, \widetilde{\widetilde{\mu}}_T^{(s)})'$, $\boldsymbol{\Phi}_{N^*\times T}^{(s)} = (\boldsymbol{\Phi}_1^{(s)\prime}, \ldots, \boldsymbol{\Phi}_T^{(s)\prime})'$, $\mathbf{A}^{(s)*} = \mathbf{I} - \mathbf{P}^{(s)*}$,

$$\boldsymbol{\Pi}_{T\times N^*}^{(s)} = \begin{bmatrix} \boldsymbol{\pi}_1^{(s)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\pi}_2^{(s)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\pi}_T^{(s)} \end{bmatrix},$$

and $\mathbf{Q}^{(s)*} = \boldsymbol{\Pi}^{(s)}\mathbf{D}^{(s)*}\boldsymbol{\Phi}^{(s)}\mathbf{A}^{(s)*}$, where $\mathbf{D}^{(s)*}$ has the same definition as $\mathbf{D}^{(s)}$ in Section 2.2 except that $\mathbf{P}_{tt}^{(s)}$ is replaced by $\mathbf{P}_{tt}^{(s)*}$. Then from (7), we obtain

$$\mathbf{Y} - \widetilde{\widetilde{\boldsymbol{\mu}}}^{(s)} = (\mathbf{A}^{(s)*} + \mathbf{Q}^{(s)*})\mathbf{Y}.$$

Thus, the criterion LsoCV($\mathbf{w}$) in (4) becomes

$$\text{LsoCV}^*(\mathbf{w}) = \mathbf{Y}'(\mathbf{A}^*(\mathbf{w}) + \mathbf{Q}^*(\mathbf{w}))'(\mathbf{A}^*(\mathbf{w}) + \mathbf{Q}^*(\mathbf{w}))\mathbf{Y},$$

where $\mathbf{Q}^*(\mathbf{w}) = \sum_{s=1}^S w_s\mathbf{Q}^{(s)*}$ and $\mathbf{A}^*(\mathbf{w}) = \sum_{s=1}^S w_s\mathbf{A}^{(s)*}$. The selected weight is given by $\widehat{\mathbf{w}}^* = \arg\min_{\mathbf{w}\in\mathcal{H}}\text{LsoCV}^*(\mathbf{w})$. In the following, we assume S is bounded and k_s is unrelated with T, where k_s is the number of columns of $\mathbf{X}^{(s)}$. Let $R_T^*(\mathbf{w}) = \|\mathbf{A}^*(\mathbf{w})\boldsymbol{\mu}\|^2 + \sigma^2 tr((\mathbf{P}^*(\mathbf{w}))^2)$, and $\xi_T^* = \inf_{\mathbf{w}\in\mathcal{H}} R_T^*(\mathbf{w})$. To show the asymptotic optimality, we further assume the following regularity conditions. All limiting processes discussed in this section are with respect to $T \to \infty$.

Condition 8. (**i**) $\{y_t, \varepsilon_t\}$ is α-mixing with size $-\gamma/(\gamma-2)$, where $\gamma > 2$.
(**ii**) $E(\varepsilon_t|\mathbf{X}_t) = 0$.
(**iii**) $E|x_{tj}^{(s)}\varepsilon_t|^\gamma \le \Delta_1 < \infty$ holds uniformly for $j = 1, \ldots, k_s$, $s = 1, \ldots, S$ and all t, where $x_{tj}^{(s)}$ is the jth element of $\mathbf{X}_t^{(s)}$ and Δ_1 is a positive constant.
(**iv**) (**a**) There exists $\delta > 0$ such that $E|x_{tj}^{(s)2}|^{\gamma/2+\delta} \le \Delta_2 < \infty$ holds uniformly for $j = 1, \ldots, k_s$, $s = 1, \ldots, S$ and all t, where Δ_2 is a positive constant.

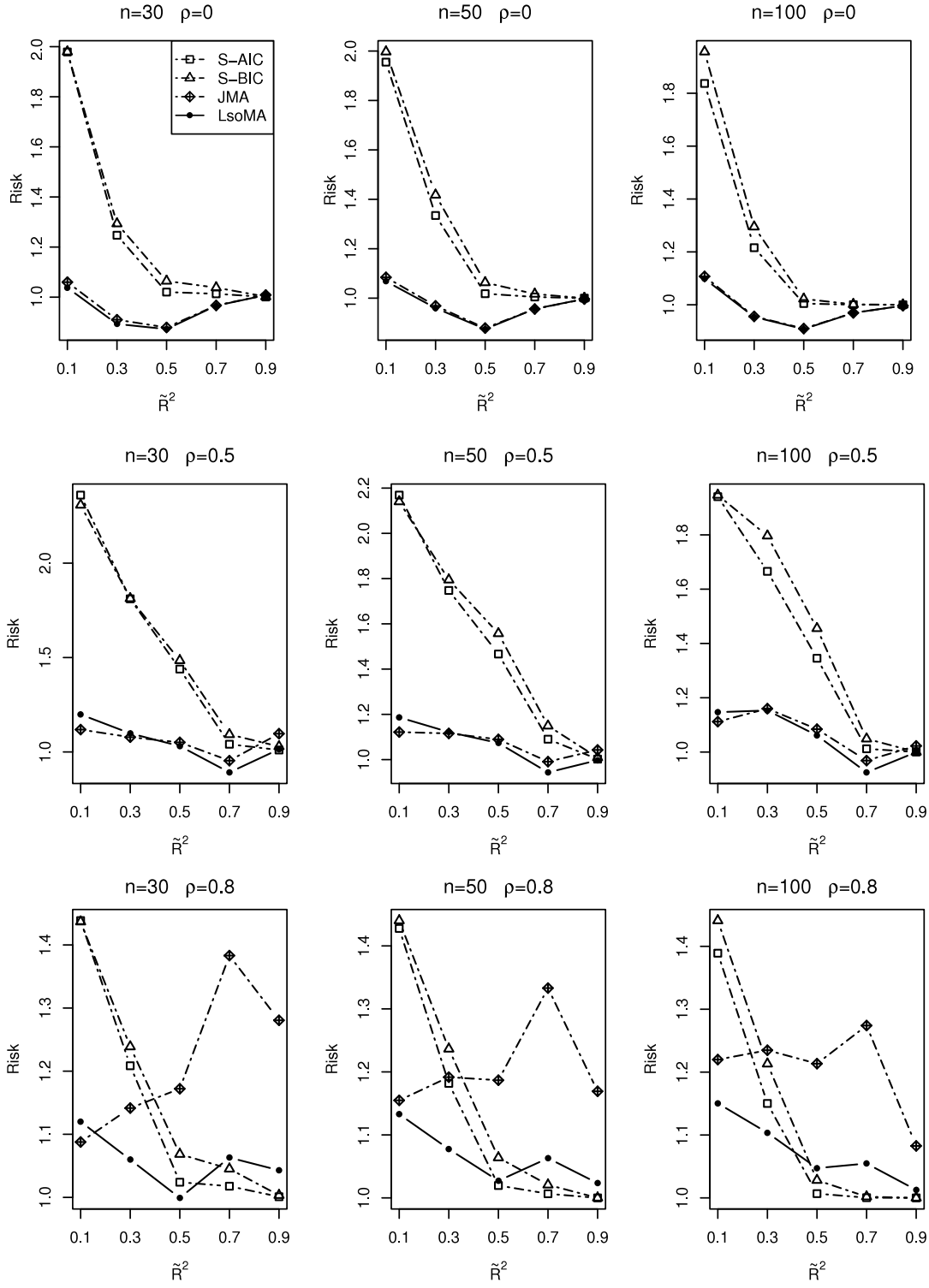

Fig. 1. Simulation results for averaging and selecting longitudinal data models.

(**b**) $\mathbf{H}_T^{(s)} \equiv E(\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)}/T)$ is uniformly positive definite for T, with $s = 1, \ldots, S$.

(**v**) $\mathbf{G}_T^{(s)} \equiv Var(\sum_{t=1}^{T} \mathbf{X}_t^{(s)}\varepsilon_t/\sqrt{T})$ is uniformly positive definite for T, with $s = 1, \ldots, S$.

Condition 9. $\sqrt{T}\xi_T^{*-1} \xrightarrow{p} 0$.

Condition 8 is a common condition to use the central limit theory for dependent heterogeneously distributed observations (White, 1984). Note that we allow for heteroscedasticity in this paper and the stationarity is not assumed. Condition 9 requires that ξ_T^* increases to infinity at a rate quicker than $\sqrt{T} \to \infty$, which is implied by the condition (7) of Ando and Li (2014) and our Condition 3.

Theorem 3.1. *Under Conditions 3, 8 and 9, we have*

$$\frac{L_T(\widehat{\mathbf{w}}^*)}{\inf_{\mathbf{w}\in\mathcal{H}} L_T(\mathbf{w})} \xrightarrow{p} 1. \tag{8}$$

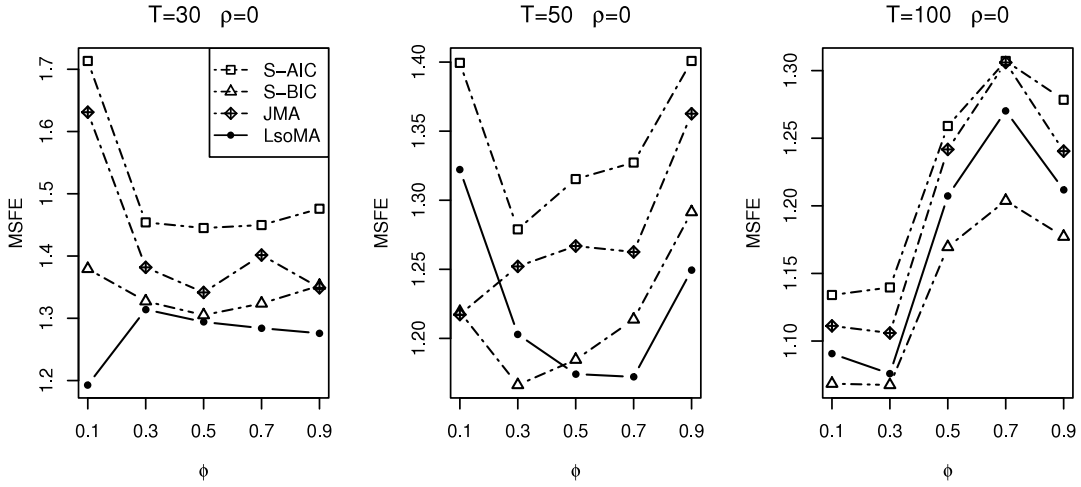

Fig. 2. Simulation results for averaging and selecting time series models with i.i.d. errors.

Proof. See Appendix.

As pointed out by a referee, Condition 8(ii) excludes the situation that ε_t is serially correlated. When considering a weakly correlated case,[4] i.e., $E(\mathbf{X}_t^{(s)}\varepsilon_t) = O(\boldsymbol{\iota}_s/\sqrt{T})$ uniformly for all t, where $\boldsymbol{\iota}_s$ is a vector of ones with dimension k_s, the asymptotic optimality shown by Theorem 3.1 still holds. The proof is similar to that of Theorem 3.1, and so only a brief demonstration is provided in the Appendix. The detailed proof is available upon request from the authors. For more general error correlation structures, we are developing methods based on the instrumental variable estimator, which is out of the scope of the current paper.

4. Simulation study

4.1. Averaging for longitudinal data models

The data generating process is as follows:

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \sum_{k=1}^{M} x_{ij,0k}\beta_{0k} + \sqrt{c_2}f(x_{ij,1}) + \varepsilon_{ij},$$
$$i = 1, \ldots, n, j = 1, \ldots, m,$$

where $\beta_{0k} = \sqrt{c_1}k^{-1}$, all the explanatory variables are independently drawn from $U(-2, 2)$, the nonlinear function is $f(x) = \sin(2x) + x/3 + x^2/2$, and ε_{ij}'s are generated from the AR(1) process $\varepsilon_{i,j} = \rho\varepsilon_{i,j-1} + v_{i,j}$ with $v_{i,j} \sim N(0, \sigma_i^2)$ for $j = 1, \ldots, m$ and $\rho \in \{0, 0.5, 0.8\}$. When $\rho = 0$, there does not exist any within-subject correlation. Set $\sigma_i^2 = \bar{x}_{i\cdot,1}^2$, where $\bar{x}_{i\cdot,1} = \sum_{j=1}^{m} x_{ij,1}/m$ and $m = 5$. Let c_1 and c_2 satisfy

$$Var\left(\sum_{k=1}^{M} x_{ij,0k}\beta_{0k}\right) / Var\left(\sqrt{c_2}f(x_{ij,1})\right) = 0.5, \tag{9}$$

i.e., the ratio of the variances of linear part and nonlinear part is 0.5. The population R^2 is given by $Var(\mu_{ij})/Var(y_{ij}) = Var(\mu_{ij})/(Var(\mu_{ij}) + Var(\varepsilon_{ij}))$. As $Var(\varepsilon_{ij})$ varies in i, we take average of σ_i^2 and get the approximate population $\widetilde{R}^2 = Var(\mu_{ij})/(Var(\mu_{ij}) + E(\sigma_i^2)/(1-\rho^2))$. We control c_1 such that $\widetilde{R}^2$ varies on a grid from 0.1 to 0.9. For a fixed c_1, the value of c_2 can be obtained by formula (9).

To construct the design and penalty matrices, we use the B-spline method discussed in Section 2.1 with $r = 4$, $K_n = 10$ and $q = r - 1 = 3$. It is clear that the tuning parameter $\lambda = 0$ corresponds to interpolation, and as λ increases, the estimation varies from the most complex model to the simplest model. Let $\mathcal{K} = \{0, h, h^2, \ldots, h^B\}$ be the candidate set of the tuning parameters with $h = 50^{1/B}$ and $B = round(n^{1/3})$, where $round(\cdot)$ stands for rounding. In addition, we set $\{\{x_{ij,01}\}, \{x_{ij,01}, x_{ij,02}\}, \ldots, \{x_{ij,01}, \ldots, x_{ij,0M}\}\}$ to be the candidate set of the linear variables, $M = 3$, and $n \in \{30, 50, 100\}$. Hence, the total number of the candidate models is $M(B+1) \in \{12, 15, 18\}$. We compare the proposed averaging estimator with estimators obtained by AIC, BIC, LsoCV, Smoothed AIC (S-AIC), Smoothed BIC (S-BIC) and JMA. To evaluate the performance of the estimators, we simulate 2000 replications and compute the relative risk, which is obtained by dividing the mean squared errors by the infeasible optimal risk (the risk of the best single model).

The results are shown in Fig. 1. In each panel, the relative risk is displayed on the y axis and $\widetilde{R}^2$ is displayed on the x axis. Since the MA estimator usually achieves a lower risk than the associated MS estimator, i.e., the MA estimators by S-AIC, S-BIC and LsoMA always perform better than the MS estimators by AIC, BIC and LsoCV, respectively, we show only the results of S-AIC, S-BIC, JMA and LsoMA so that we can distinguish different lines clearly. Simulation results of the MS estimators are available upon request from the authors. It can be seen from Fig. 1 that LsoMA performs much better than S-AIC and S-BIC in most cases. When $\rho = 0$, LsoMA and JMA perform closely; while for $\rho \in \{0.5, 0.8\}$, LsoMA is better than JMA, especially for large $\widetilde{R}^2$. It can also be observed that for large $\rho(=0.8)$, JMA often leads to a much larger risk than S-AIC and S-BIC.

4.2. Averaging for time series models

The data generating process is as follows:

$$y_t = f(y_{t-1}) + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \alpha_4 y_{t-4} + e_t,$$

where $t = 1, \ldots, T$, $\alpha_j = 2(-\phi)^{j-2}/(3\sum_{j'=2}^{4}\phi^{j'-2})$, e_t follows an AR(1) process $e_t = \rho e_{t-1} + v_t$ with v_t being i.i.d. $\sim N(0, 1)$, and $\rho \in \{0, 1/\sqrt{T}, 3/\sqrt{T}, 5/\sqrt{T}\}$. When we choose $\rho = 0$, it reduces to the case that e_t is i.i.d. $\sim N(0, 1)$. The other choices of ρ represent the case that e_t is weakly correlated. Note that ϕ determines the

[4] This weakly correlated case belongs to a local-to-zero asymptotic framework which is also used in other model averaging papers such as Hjort and Claeskens (2003) and Liu (2015).

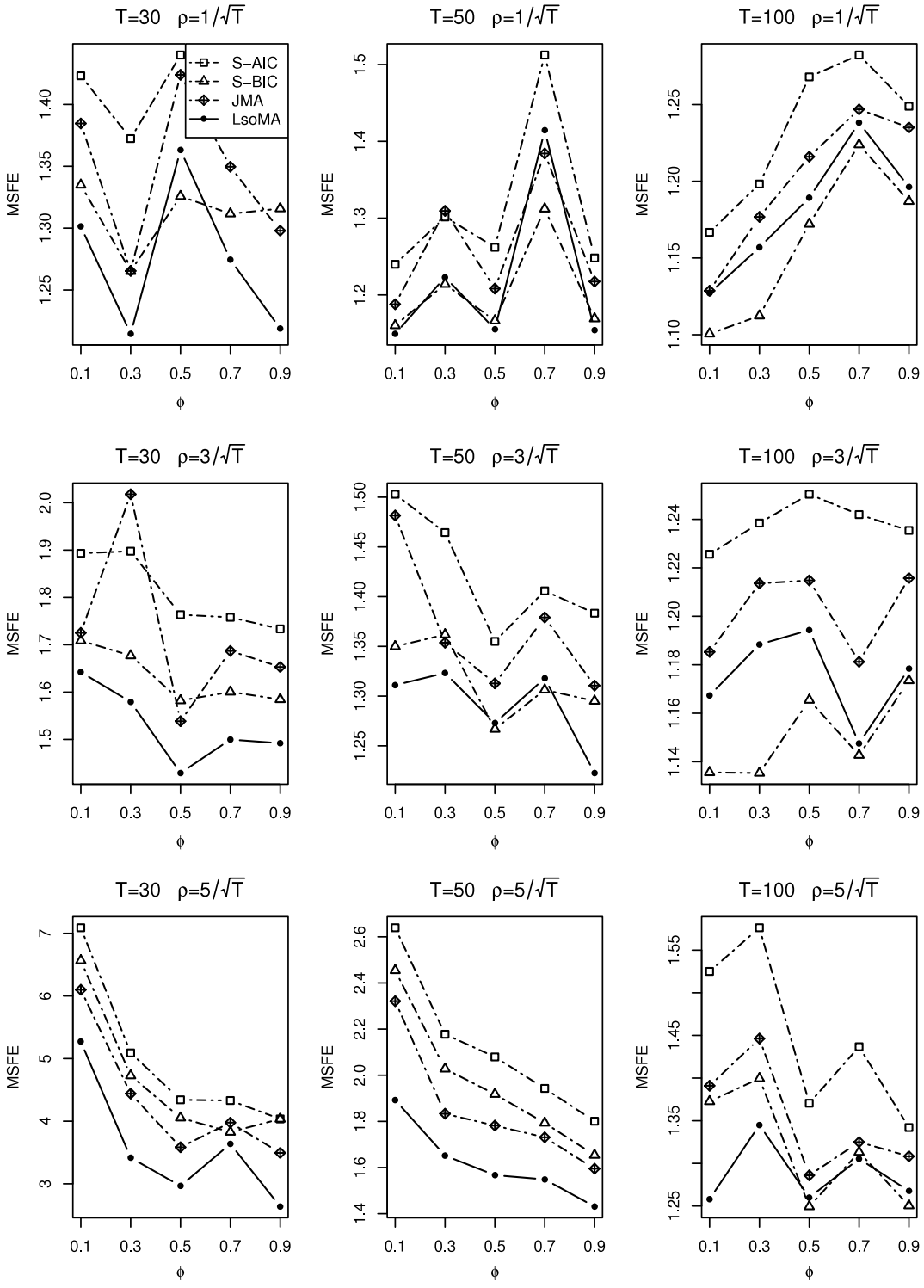

Fig. 3. Simulation results for averaging and selecting time series models with weakly correlated errors.

convergence rate of the linear coefficients. Let ϕ vary on a grid from 0.1 to 0.9. Similar to Fan and Yao (2003), we set the nonlinear function to be $f(x) = (0.316+0.982x)e^{-3.89x^2}/3$. The sth candidate model is

$$y_t = f^{(s)}(y_{t-1}) + \alpha_2^{(s)} y_{t-2} + \cdots + \alpha_{p_s+1}^{(s)} y_{t-p_s-1} + e_t.$$

We set $p_s \in \{1, 2, 3\}$. To get the design and penalty matrices, we take $r = 3$, $q = 2$ and $K_n = 10$. Similar to Section 4.1, the candidate set of λ is set to be $\{1, h, h^2, \ldots, h^B\}$, where $h = 50^{1/B}$, and $B = round(T^{1/3})$. Let $T = 30, 50, 100$.

For the sth candidate model, the one-step-ahead out-of-sample forecast of y_{T+1} is

$$\widehat{y}_{T+1}^{(s)} = \widehat{f}^{(s)}(y_T) + \widehat{\alpha}_2^{(s)} y_{T-1} + \cdots + \widehat{\alpha}_{p_s+1}^{(s)} y_{T-p_s},$$

where $\widehat{f}^{(s)}(y_T)$ is obtained from the B-spline basis function. Then the combined forecast of y_{T+1} is given by $\widehat{y}_{T+1}(w) = \sum_{s=1}^{S} w_s \widehat{y}_{T+1}^{(s)}$.

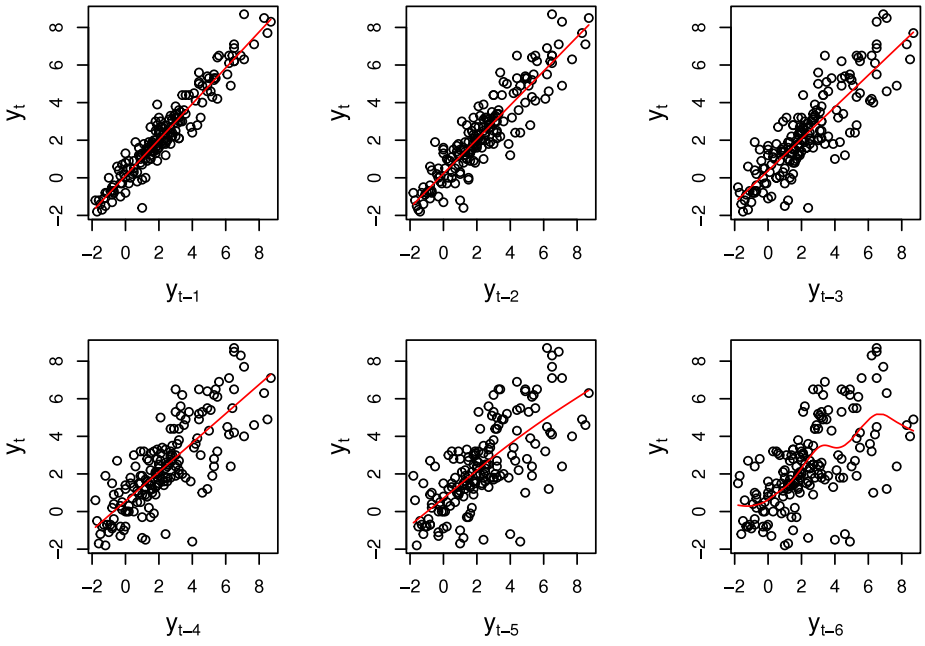

Fig. 4. The relationship between y_t and its lag variables in the empirical application on Chinese CPI.

To assess the accuracy of forecasts based on different methods, we calculate the mean squared forecast error (MSFE)

$$\sum_{d=1}^{D} \|\widehat{y}_{T+1}^{(d)}(w) - y_{T+1}^{(d)}\|^2/D,$$

where $D = 2000$ is the number of simulation trials and d denotes the dth run. We compare our proposed method with the commonly used MS and MA methods: AIC, BIC, LsoCV, S-AIC, S-BIC and JMA.

The results are depicted in Figs. 2–3. As in Simulation I, we show only the results of the MA methods. It is clear that LsoMA performs better than the other MA methods in most cases. When T is large and ρ is small (for example, $T = 100$ and $\rho = 0$), S-BIC can achieve a lower MSFE than LsoMA. For almost all the cases, S-AIC is the worst.

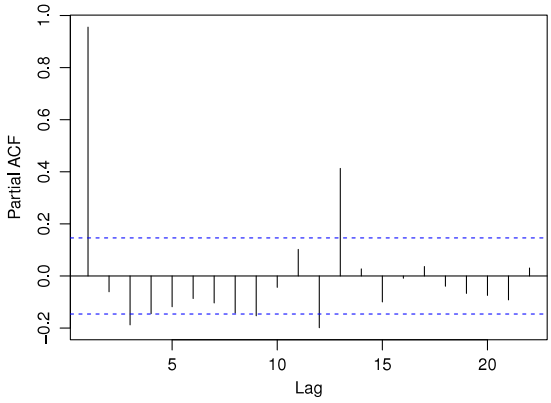

Fig. 5. PACF of Chinese CPI.

5. Empirical application

In this section, we use the MS and MA methods to analyze the Chinese Consumer Price Index (CPI) data. The series, denoted as $\{y_t\}$, is monthly from January 2000 to December 2014. We first conduct some preliminary analysis of the data. The Augmented Dickey–Fuller (ADF) test shows that the series is stationary at the significant level of 0.05. Fig. 4 describes the relationship between y_t and its lag variable y_{t-m}. When $m \leq 5$, the fitted line is linear; but when $m = 6$, the fitted line is nonlinear. Further, the fitted lines for $m = 7, \ldots, 15$ are all nonlinear, which are not shown in the paper but available upon request from the authors. Fig. 5 describes the PACFs, indicating that from the 6th to 22nd lag variables, the 13th has the largest PACF. So we set y_{t-13} as the nonlinear variable and $y_1, \ldots, y_5$ as the linear variables. Thus, the candidate models are given by

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_{p_s} y_{t-p_s} + f^{(s)}(y_{t-13}) + e_t, \quad t = 1, \ldots, T \tag{10}$$

with $p_s = 1, \ldots, 5$, where $\alpha_0, \ldots, \alpha_{p_s}$ are the coefficients, $f^{(s)}(\cdot)$ is the smooth function, and e_t is the error term.

We compare model selection and averaging methods: AIC, BIC, LsoCV, S-AIC, S-BIC, JMA, and LsoMA. We examine these methods by using one-step-ahead out-of-sample forecast. When forecasting y_t, the T observations before y_t are used. We set $T = 72$ (6 years), 108 (9 years), and 144 (12 years). In addition, following a comment by a referee, we also include the results obtained by averaging linear models. Specifically, we use the LsoMA method to average forecasts from the linear models $y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_{p_s} y_{t-p_s} + \alpha_{13} y_{t-13} + e_t$ with $p_s = 1, \ldots, 5$. Accordingly, the LsoMA method is labeled as LsoMA_L.

The means and medians of the squared forecast errors with their standard errors are listed in Table 1. It is readily seen that as T increases, all methods have better performances. The difference among the medians of all methods is clearer than that among their means. Depending on means, LsoMA is superior to all other methods for $T = 72$ and $T = 108$. When $T = 144$, LsoCV performs the best and LsoMA performs the second best, but their difference is very minor. Depending on medians, LsoMA outperforms the other methods in all cases.

We also draw the predicted lines based on different methods. Fig. 6 shows the comparison results between LsoMA and JMA. The in-sample predictions are drawn on the left-hand side of the vertical dotted line, while the out-of-sample predictions are drawn on the other side. It can be observed that LsoMA performs better

Table 1
Means and medians of the squared forecast errors in forecasting Chinese CPI (standard errors are in brackets).

$T = 72$	AIC	BIC	LsoCV	S-AIC	S-BIC	JMA	LsoMA	LsoMA_L
Mean	0.4300 *(0.0689)*	0.4092 *(0.0717)*	0.4114 *(0.0917)*	0.4220 *(0.0787)*	0.4090 *(0.0747)*	0.3976 *(0.0716)*	0.3966 *(0.0827)*	0.4008 *(0.0724)*
Median	0.2629 *(0.0863)*	0.1776 *(0.0898)*	0.1766 *(0.1149)*	0.2241 *(0.0986)*	0.1848 *(0.0937)*	0.2193 *(0.0898)*	0.1681 *(0.1036)*	0.1979 *(0.0908)*
$T = 108$								
Mean	0.3858 *(0.0612)*	0.3737 *(0.0621)*	0.3598 *(0.0824)*	0.3756 *(0.0724)*	0.3723 *(0.0706)*	0.3738 *(0.0636)*	0.3531 *(0.0737)*	0.3707 *(0.0651)*
Median	0.1793 *(0.0767)*	0.1500 *(0.0779)*	0.1294 *(0.1033)*	0.1496 *(0.0908)*	0.1326 *(0.0885)*	0.1568 *(0.0797)*	0.1271 *(0.0924)*	0.1394 *(0.0815)*
$T = 144$								
Mean	0.2686 *(0.0300)*	0.2499 *(0.0305)*	0.2471 *(0.0302)*	0.2561 *(0.0303)*	0.2525 *(0.0310)*	0.2568 *(0.0298)*	0.2475 *(0.0313)*	0.2565 *(0.0319)*
Median	0.1499 *(0.0376)*	0.1395 *(0.0383)*	0.1403 *(0.0378)*	0.1506 *(0.0379)*	0.1363 *(0.0389)*	0.1407 *(0.0374)*	0.1261 *(0.0392)*	0.1411 *(0.0400)*

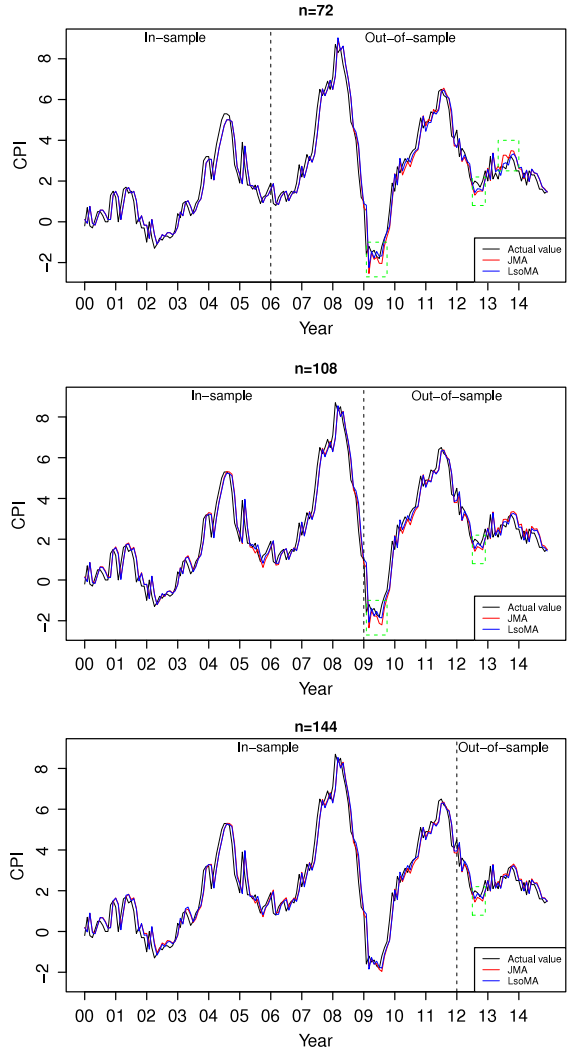

Fig. 6. Predicted Chinese CPI values by JMA and LsoMA.

than JMA when the CPI fluctuates sharply. For example, with $T = 72$ and $T = 108$, the biases of JMA are larger than those of LsoMA during 2009, when the CPI decreases to the bottom due to the US subprime mortgage crisis. Another evidence of this phenomenon appears in the second half of 2012. In other periods, the prediction lines of JMA and LsoMA are nearly the same. The comparison results between LsoMA and other methods are similar, and are available upon request from the authors.

6. Conclusion

This paper develops a leave-subject-out model averaging (LsoMA) method for semiparametric longitudinal data model and serially dependent time series model. The method can average estimators based on different regressors and/or different tuning parameters. The asymptotic optimality of the LsoMA estimators has been proved. Simulation studies show the promise of the LsoMA method in both longitudinal and time series data. The empirical analysis on Chinese CPI data also shows superiority of the proposed method.

Our method is based on the given set $\{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(K)}\}$. Xu and Huang (2012) proposed an efficient Newton type algorithm to compute the optimal $\boldsymbol{\lambda}$. One can use $\boldsymbol{\lambda}$'s generated from the algorithm procedure as the candidate tuning parameters. However, the theoretical justification for using such tuning parameters needs further investigation. Additionally, developing the asymptotic distributions of our MA estimators is also a challenging and important topic, and this warrants our future research.

Acknowledgments

We thank the editor Oliver Linton, the associate editor, the two anonymous referees and Dr. Ganggang Xu for their constructive comments and suggestions which greatly improved the original manuscript. Zhang's work was partially supported by the National Natural Science Foundation of China (Grant nos. 71522004 and 11471324) and a special talent grant from AMSS, CAS. Zou's work was partially supported by the National Natural Science Foundation of China (Grant nos. 11331011 and 11271355) and a grant from the Beijing High-level Talents Program.

Appendix

Lemma 1. *We have the following inequalities,*

(a) $\bar{\sigma}(\widetilde{\mathbf{P}}^{(s)}) \leq 1$,

(b) $\bar{\sigma}(\mathbf{I} - \widetilde{\mathbf{P}}^{(s)}) \leq 1$.

If Condition 7 holds, then

(c) $\bar{\sigma}(\mathbf{P}^{(s)}) \leq \sqrt{C_3}$,

(d) $tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}) \leq C_3 k_s$.

Proof. First note that for any positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A} \leq \mathbf{B}$, then $\bar{\sigma}(\mathbf{A}) \leq \bar{\sigma}(\mathbf{B})$. Since $\widetilde{\mathbf{P}}^{(s)}$ is obviously positive

definite and

$$\widetilde{\mathbf{P}}^{(s)} = \mathbf{V}_{(s)}^{-1/2}\mathbf{X}^{(s)}\left(\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1}\mathbf{X}^{(s)} + \sum_{l=1}^{L}\lambda_l^{(s)}\mathbf{F}_l^{(s)}\right)^{-1}\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1/2} \le \mathbf{V}_{(s)}^{-1/2}\mathbf{X}^{(s)}(\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1}\mathbf{X}^{(s)})^{-1}\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1/2},$$

we have

$$\bar{\sigma}(\widetilde{\mathbf{P}}^{(s)}) \le \bar{\sigma}(\mathbf{V}_{(s)}^{-1/2}\mathbf{X}^{(s)}(\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1}\mathbf{X}^{(s)})^{-1}\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1/2}) = 1$$

and

$$\bar{\sigma}(\mathbf{I} - \widetilde{\mathbf{P}}^{(s)}) \le \bar{\sigma}(\mathbf{I}) = 1.$$

Thus (a) and (b) are valid.

Since $\mathbf{X}^{(s)}$ has full column rank k_s, we have

$$tr(\mathbf{P}^{(s)}) = tr(\widetilde{\mathbf{P}}^{(s)}) \le tr(\mathbf{V}_{(s)}^{-1/2}\mathbf{X}^{(s)}(\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1}\mathbf{X}^{(s)})^{-1}\mathbf{X}^{(s)\prime}\mathbf{V}_{(s)}^{-1/2}) = k_s.$$

From (a) and Condition 7, it can be shown that

$$\bar{\sigma}(\mathbf{P}^{(s)}) = \bar{\sigma}(\mathbf{V}_{(s)}^{1/2}\widetilde{\mathbf{P}}^{(s)}\mathbf{V}_{(s)}^{-1/2}) \le \bar{\sigma}(\mathbf{V}_{(s)}^{1/2})\bar{\sigma}(\mathbf{V}_{(s)}^{-1/2})\bar{\sigma}(\widetilde{\mathbf{P}}^{(s)}) \le \sqrt{C_3}$$

and

$$tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}) = tr(\mathbf{V}_{(s)}^{-1/2}\widetilde{\mathbf{P}}^{(s)}\mathbf{V}_{(s)}\widetilde{\mathbf{P}}^{(s)}\mathbf{V}_{(s)}^{-1/2}) \le \bar{\sigma}(\mathbf{V}_{(s)}^{-1})\bar{\sigma}(\mathbf{V}_{(s)})\bar{\sigma}(\widetilde{\mathbf{P}}^{(s)})tr(\widetilde{\mathbf{P}}^{(s)}) \le C_3k_s.$$

So (c) and (d) are true. The proof of Lemma 1 is completed.

Let $\mathbf{M}(\mathbf{w}) = \mathbf{A}'(\mathbf{w})\mathbf{Q}(\mathbf{w}) + \mathbf{Q}'(\mathbf{w})\mathbf{A}(\mathbf{w}) + \mathbf{Q}'(\mathbf{w})\mathbf{Q}(\mathbf{w})$. We have the following lemma.

Lemma 2. *If Conditions 2(i) and 7 hold, then*

$$\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{M}(\mathbf{w})) = O_p(k_{S^*}/n).$$

Proof. Using Lemma 1(a), Condition 7, and some simple inequalities on singular value, we have

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{A}(\mathbf{w})) &\le \sup_{\mathbf{w}\in\mathcal{H}}\sum_{s=1}^{S}w_s\bar{\sigma}(\mathbf{I}-\mathbf{P}^{(s)}) \\ &= \sup_{\mathbf{w}\in\mathcal{H}}\sum_{s=1}^{S}w_s\bar{\sigma}(\mathbf{V}_{(s)}^{1/2}(\mathbf{I}-\widetilde{\mathbf{P}}^{(s)})\mathbf{V}_{(s)}^{-1/2}) \\ &\le \max_s\bar{\sigma}(\mathbf{V}_{(s)}^{1/2})\bar{\sigma}(\mathbf{V}_{(s)}^{-1/2}) \\ &= \max_s c(\mathbf{V}_{(s)}^{1/2}) \le \sqrt{C_3}. \end{aligned} \tag{A.1}$$

Let $\widetilde{\mathbf{D}}^{(s)} = \mathbf{V}_{(s)}^{-1/2}\mathbf{D}^{(s)}\mathbf{V}_{(s)}^{1/2}$ and $\widetilde{\mathbf{D}}_{ii}^{(s)}$ be the ith diagonal block of $\widetilde{\mathbf{D}}^{(s)}$. Noting that

$$tr(\widetilde{\mathbf{P}}_{ii}^{(s)}) = tr(\mathbf{P}_{ii}^{(s)}\mathbf{V}_{(s)i}^{1/2}\mathbf{V}_{(s)i}^{-1/2}) = tr(\mathbf{P}_{ii}^{(s)}) \tag{A.2}$$

and $\widetilde{\mathbf{D}}^{(s)}$ is a positive semi-definite and block diagonal matrix, we obtain

$$\begin{aligned}\bar{\sigma}(\widetilde{\mathbf{D}}^{(s)}) &\le \max_i tr\widetilde{\mathbf{D}}_{ii}^{(s)} = \max_i tr\left(\sum_{k=1}^{\infty}(\widetilde{\mathbf{P}}_{ii}^{(s)})^k\right) = \max_i\sum_{k=1}^{\infty}tr(\widetilde{\mathbf{P}}_{ii}^{(s)})^k \\ &\le \max_i\sum_{k=1}^{\infty}(tr\widetilde{\mathbf{P}}_{ii}^{(s)})^k = \max_i\frac{tr\widetilde{\mathbf{P}}_{ii}^{(s)}}{1-tr\widetilde{\mathbf{P}}_{ii}^{(s)}} = \max_i\frac{tr\mathbf{P}_{ii}^{(s)}}{1-tr\mathbf{P}_{ii}^{(s)}}.\end{aligned}$$

Then by Lemma 1(b) and Conditions 2(i) and 7, it can be seen that

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{Q}(\mathbf{w})) &\le \sup_{\mathbf{w}\in\mathcal{H}}\sum_{s=1}^{S}w_s\bar{\sigma}(\mathbf{D}^{(s)}(\mathbf{I}-\mathbf{P}^{(s)})) \\ &= \sup_{\mathbf{w}\in\mathcal{H}}\sum_{s=1}^{S}w_s\bar{\sigma}(\mathbf{V}_{(s)}^{1/2}\widetilde{\mathbf{D}}^{(s)}(\mathbf{I}-\widetilde{\mathbf{P}}^{(s)})\mathbf{V}_{(s)}^{-1/2}) \\ &= \max_s\bar{\sigma}(\mathbf{V}_{(s)}^{1/2})\bar{\sigma}(\mathbf{V}_{(s)}^{-1/2})\bar{\sigma}(\widetilde{\mathbf{D}}^{(s)})\bar{\sigma}(\mathbf{I}-\widetilde{\mathbf{P}}^{(s)}) \\ &\le \sqrt{C_3}\max_i\max_s\frac{tr\mathbf{P}_{ii}^{(s)}}{1-tr\mathbf{P}_{ii}^{(s)}} = O_p(k_{S^*}/n). \end{aligned} \tag{A.3}$$

From (A.1) to (A.3), we have

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{M}(\mathbf{w})) &\le 2\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{A}'(\mathbf{w})\mathbf{Q}(\mathbf{w})) + \sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{Q}'(\mathbf{w})\mathbf{Q}(\mathbf{w})) \\ &\le 2\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{A}(\mathbf{w}))\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{Q}(\mathbf{w})) + \left(\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{Q}(\mathbf{w}))\right)^2 \\ &= O_p(k_{S^*}/n). \end{aligned} \tag{A.4}$$

The proof of Lemma 2 is completed.

Proof of Theorem 2.1. Note that

$$\begin{aligned}\mathrm{LsoCV}(\mathbf{w}) &= \mathbf{Y}'\mathbf{A}'(\mathbf{w})\mathbf{A}(\mathbf{w})\mathbf{Y} + \mathbf{Y}'\mathbf{M}(\mathbf{w})\mathbf{Y} \\ &= L_n(\mathbf{w}) + \boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\mu} + \boldsymbol{\varepsilon}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon} + 2\boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon} \\ &\quad + 2\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}.\end{aligned}$$

Since $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ is unrelated to $\mathbf{w}$, minimizing $\mathrm{LsoCV}(\mathbf{w})$ is equivalent to minimizing $\mathrm{LsoCV}(\mathbf{w}) - \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. So we need only to verify that

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\mu}|/R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{A.5}$$

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{A.6}$$

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{A.7}$$

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{A.8}$$

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) \xrightarrow{p} 0, \tag{A.9}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right| \xrightarrow{p} 0. \tag{A.10}$$

From Lemma 2 and Conditions 3 and 6, it can be seen that

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\mu}|/R_n(\mathbf{w}) \le \|\boldsymbol{\mu}\|^2\xi_n^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{M}(\mathbf{w})) = \|\boldsymbol{\mu}\|^2\xi_n^{-1}O_p(k_{S^*}/n) = o_p(1).$$

So (A.5) is true.

By Condition 5, we have $E\{(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})^{1/2}\} \le \{E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})\}^{1/2} \le \sqrt{N\bar{\sigma}(\boldsymbol{\Sigma})} = O(\sqrt{N})$, and thus, $(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})^{1/2} = O_p(\sqrt{N})$ and $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = O_p(N)$. Consequently, from Lemma 2 and Conditions 3 and 6, we obtain

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) &\le \xi_n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{M}(\mathbf{w})) \\ &= \xi_n^{-1}O_p(k_{S^*}/n)O_p(N) = o_p(1)\end{aligned}$$

and

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) &\le \xi_n^{-1}\sup_{\mathbf{w}\in\mathcal{H}}(\boldsymbol{\varepsilon}'\mathbf{M}(\mathbf{w})\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{M}(\mathbf{w})\boldsymbol{\varepsilon})^{1/2}\\ &\le \xi_n^{-1}\|\boldsymbol{\mu}\|(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})^{1/2}\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{M}(\mathbf{w}))\\ &= \xi_n^{-1}\frac{N}{n}O_p(k_{S^*}) = o_p(1).\end{aligned}$$

So (A.6) and (A.7) are correct.

To show (A.8) and (A.9), we first assume that $\mathbf{X}$ is non-random. Let $\mathbf{B}^{(s)} = \boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(s)}\boldsymbol{\Sigma}^{1/2}$ and $\mathbf{B}_{ii}^{(s)*} = \boldsymbol{\Sigma}_i^{1/2}[\mathbf{V}_{(s)i}^{1/2}\widetilde{\mathbf{P}}_{ii}^{(s)}\mathbf{V}_{(s)i}^{1/2}/\bar{\sigma}(\mathbf{V}_{(s)i}) + \bar{\sigma}(\mathbf{V}_{(s)i})\mathbf{V}_{(s)i}^{-1/2}\widetilde{\mathbf{P}}_{ii}^{(s)}\mathbf{V}_{(s)i}^{-1/2}]\boldsymbol{\Sigma}_i^{1/2}/2$. According to (1), we have $tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}\boldsymbol{\Sigma}) \le R_n(\mathbf{w}_s^0)$, and so

$$\begin{aligned}tr(\mathbf{B}^{(s)}\mathbf{B}^{(s)\prime}) &= tr(\boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(s)}\boldsymbol{\Sigma}\mathbf{P}^{(s)\prime}\boldsymbol{\Sigma}^{1/2}) \le \bar{\sigma}(\boldsymbol{\Sigma})tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}\boldsymbol{\Sigma})\\ &\le \bar{\sigma}(\boldsymbol{\Sigma})R_n(\mathbf{w}_s^0).\end{aligned}\tag{A.11}$$

From (A.2) and Conditions 2(i), 5 and 7, it can be seen that

$$\begin{aligned}tr(\mathbf{B}_{ii}^{(s)*}) &= \frac{1}{2}tr(\mathbf{V}_{(s)i}^{1/2}\widetilde{\mathbf{P}}_{ii}^{(s)}\mathbf{V}_{(s)i}^{1/2}\boldsymbol{\Sigma}_i)/\bar{\sigma}(\mathbf{V}_{(s)i})\\ &\quad+\frac{1}{2}\bar{\sigma}(\mathbf{V}_{(s)i})tr(\mathbf{V}_{(s)i}^{-1/2}\widetilde{\mathbf{P}}_{ii}^{(s)}\mathbf{V}_{(s)i}^{-1/2}\boldsymbol{\Sigma}_i)\\ &\le \frac{1}{2}\bar{\sigma}(\boldsymbol{\Sigma}_i)\bar{\sigma}(\mathbf{V}_{(s)i})tr(\widetilde{\mathbf{P}}_{ii}^{(s)})/\bar{\sigma}(\mathbf{V}_{(s)i})\\ &\quad+\frac{1}{2}\bar{\sigma}(\boldsymbol{\Sigma}_i)\bar{\sigma}(\mathbf{V}_{(s)i})\bar{\sigma}(\mathbf{V}_{(s)i}^{-1})tr(\widetilde{\mathbf{P}}_{ii}^{(s)})\\ &= O(k_{S^*}/n)\end{aligned}$$

uniformly in i and s. So

$$\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^{(s)*})\}^2 = O(k_{S^*}^2/n)\tag{A.12}$$

uniformly in s. By Lemma A.4 of Xu and Huang (2012), Condition 1, (A.11) and (A.12), we have

$$\begin{aligned}Var(\boldsymbol{\varepsilon}'\mathbf{P}^{(s)}\boldsymbol{\varepsilon}) &\le 2tr(\mathbf{B}^{(s)}\mathbf{B}^{(s)\prime}) + C_1\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^{(s)*})\}^2\\ &= O(\bar{\sigma}(\boldsymbol{\Sigma})R_n(\mathbf{w}_s^0)) + O(k_{S^*}^2/n).\end{aligned}$$

Thus, from the Chebyshev's inequality and Conditions 3–6, we see that for any $\delta > 0$,

$$\begin{aligned}&\Pr\left\{\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon} - tr(\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})|/R_n(\mathbf{w}) > \delta\right\}\\ &\le \sum_{s=1}^{S}\Pr\{|\boldsymbol{\varepsilon}'\mathbf{P}^{(s)}\boldsymbol{\varepsilon} - tr(\mathbf{P}^{(s)}\boldsymbol{\Sigma})| > \delta\xi_n\}\\ &\le \delta^{-2}\xi_n^{-2}\sum_{s=1}^{S}Var(\boldsymbol{\varepsilon}'\mathbf{P}^{(s)}\boldsymbol{\varepsilon})\\ &= O\left(\bar{\sigma}(\boldsymbol{\Sigma})\delta^{-2}\xi_n^{-2}\sum_{s=1}^{S}R_n(\mathbf{w}_s^0)\right) + O(Sn^{-1}k_{S^*}^2\xi_n^{-2})\\ &= O\left(\bar{\sigma}(\boldsymbol{\Sigma})\delta^{-2}\xi_n^{-2}\sum_{s=1}^{S}R_n(\mathbf{w}_s^0)\right)\\ &\quad+O((Nk_{S^*}Sn^{-1}\xi_n^{-1})k_{S^*}N^{-1}\xi_n^{-1}) \to 0.\end{aligned}\tag{A.13}$$

On the other hand, from Condition 5 and Lemma 1(c), we obtain

$$\begin{aligned}|tr(\mathbf{P}^{(s)}\boldsymbol{\Sigma})| &= \frac{1}{2}|tr(\mathbf{P}^{(s)}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{P}^{(s)'})|\\ &\le \frac{1}{2}\bar{\sigma}(\mathbf{P}^{(s)}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{P}^{(s)'})\text{rank}(\mathbf{P}^{(s)}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{P}^{(s)'})\\ &\le 2\bar{\sigma}(\mathbf{P}^{(s)})\bar{\sigma}(\boldsymbol{\Sigma})\text{rank}(\mathbf{P}^{(s)}\boldsymbol{\Sigma}) \le 2\sqrt{C_3}C_2k_s,\end{aligned}$$

and so

$$|tr(\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})| \le \sum_{s=1}^{S}w_s|tr(\mathbf{P}^{(s)}\boldsymbol{\Sigma})| \le 2\sqrt{C_3}C_2k_{S^*}.$$

Hence, by Condition 6, we have

$$\sup_{\mathbf{w}\in\mathcal{H}}|tr(\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})|/R_n(\mathbf{w}) \le 2\sqrt{C_3}C_2k_{S^*}\xi_n^{-1} = o(1).\tag{A.14}$$

(A.8) is now clear from (A.13) to (A.14).

From the Chebyshev's inequality, (1), and Conditions 4–5, for any $\delta > 0$, we have

$$\begin{aligned}&\Pr\left\{\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) > \delta\right\} \le \Pr\left\{\max_{1\le s\le S}|\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w}_s^0)\boldsymbol{\varepsilon}| > \delta\xi_n\right\}\\ &\le \sum_{s=1}^{S}\Pr\{|\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w}_s^0)\boldsymbol{\varepsilon}| > \delta\xi_n\} \le \delta^{-2}\xi_n^{-2}\sum_{s=1}^{S}E(\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w}_s^0)\boldsymbol{\varepsilon})^2\\ &= \delta^{-2}\xi_n^{-2}\sum_{s=1}^{S}tr(\boldsymbol{\Sigma}\mathbf{A}(\mathbf{w}_s^0)\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w}_s^0))\\ &\le \delta^{-2}\xi_n^{-2}\bar{\sigma}(\boldsymbol{\Sigma})\sum_{s=1}^{S}\|\mathbf{A}(\mathbf{w}_s^0)\boldsymbol{\mu}\|^2\\ &\le \delta^{-2}\xi_n^{-2}C_2\sum_{s=1}^{S}R_n(\mathbf{w}_s^0) \to 0.\end{aligned}\tag{A.15}$$

So (A.9) holds. When $\mathbf{X}$ is random, the above proof steps and the dominated convergence theorem imply that (A.8) and (A.9) are both correct.

Since

$$\begin{aligned}&L_n(\mathbf{w}) - R_n(\mathbf{w})\\ &= \boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon} - 2\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon} - tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma}),\end{aligned}$$

(A.10) is implied by

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon} - tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})|/R_n(\mathbf{w}) \xrightarrow{p} 0\tag{A.16}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) \xrightarrow{p} 0.\tag{A.17}$$

Similar to the proofs of (A.8) and (A.9), for showing (A.16) and (A.17), we can consider only the case where $\mathbf{X}$ is non-random. Let $\mathbf{B}^{(st)} = 2\boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}\boldsymbol{\Sigma}^{1/2}$, $\mathbf{B}^{(st)*} = \boldsymbol{\Sigma}^{1/2}(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)} + \mathbf{P}^{(t)\prime}\mathbf{P}^{(t)})\boldsymbol{\Sigma}^{1/2}$ and $\mathbf{B}_{ii}^{(st)*} = \mathbf{L}_i\mathbf{B}^{(st)*}\mathbf{L}_i'$. From Lemma 1(c)–(d) and Condition 5, we have

$$\begin{aligned}tr(\mathbf{B}^{(st)}\mathbf{B}^{(st)\prime}) &= 4tr(\boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}\boldsymbol{\Sigma}\mathbf{P}^{(t)\prime}\mathbf{P}^{(s)}\boldsymbol{\Sigma}^{1/2})\\ &\le 4\bar{\sigma}(\boldsymbol{\Sigma})tr(\mathbf{P}^{(s)}\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}\boldsymbol{\Sigma}\mathbf{P}^{(t)\prime})\\ &\le 4\bar{\sigma}(\boldsymbol{\Sigma})\bar{\sigma}^2(\mathbf{P}^{(s)})tr(\mathbf{P}^{(t)\prime}\mathbf{P}^{(t)}\boldsymbol{\Sigma})\\ &\le 4\bar{\sigma}^2(\boldsymbol{\Sigma})\bar{\sigma}^2(\mathbf{P}^{(s)})tr(\mathbf{P}^{(t)\prime}\mathbf{P}^{(t)})\\ &\le 4C_2^2C_3^2k_t\end{aligned}\tag{A.18}$$

and

$$\begin{aligned}\sum_{i=1}^{n} tr(\mathbf{B}_{ii}^{(st)*}) &= tr(\mathbf{B}^{(st)*}) \\ &= tr(\boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}\boldsymbol{\Sigma}^{1/2}) + tr(\boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(t)\prime}\mathbf{P}^{(t)}\boldsymbol{\Sigma}^{1/2}) \\ &\le \bar{\sigma}(\boldsymbol{\Sigma})tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}) + \bar{\sigma}(\boldsymbol{\Sigma})tr(\mathbf{P}^{(t)\prime}\mathbf{P}^{(t)}) \\ &\le C_2C_3(k_s + k_t). \end{aligned} \tag{A.19}$$

Using Conditions 2(ii), 5 and 7, we obtain

$$\begin{aligned} &tr(\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i') \\ &= tr(\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{V}_{(s)}^{-1/2}\widetilde{\mathbf{P}}^{(s)}\mathbf{V}_{(s)}\widetilde{\mathbf{P}}^{(s)}\mathbf{V}_{(s)}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i') \\ &\le \bar{\sigma}(\mathbf{V}_{(s)})tr(\widetilde{\mathbf{P}}^{(s)}\mathbf{V}_{(s)}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i'\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{V}_{(s)}^{-1/2}\widetilde{\mathbf{P}}^{(s)}) \\ &= \bar{\sigma}(\mathbf{V}_{(s)})tr((\widetilde{\mathbf{P}}^{(s)})_{ii}^2\mathbf{V}_{(s)i}^{-1/2}\boldsymbol{\Sigma}_i\mathbf{V}_{(s)i}^{-1/2}) \\ &\le \bar{\sigma}(\mathbf{V}_{(s)})\bar{\sigma}(\mathbf{V}_{(s)}^{-1})\bar{\sigma}(\boldsymbol{\Sigma})tr((\widetilde{\mathbf{P}}^{(s)})_{ii}^2) = o(1) \end{aligned} \tag{A.20}$$

uniformly in i and s. Thus, by (A.19) and (A.20), we have

$$\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^{(st)*})\}^2 \le \max_i tr(\mathbf{B}_{ii}^{(st)*})\sum_{i=1}^{n} tr(\mathbf{B}_{ii}^{(st)*}) = o(k_{S^*}) \tag{A.21}$$

uniformly in s and t. From Lemma A.4 of Xu and Huang (2012), Condition 1, (A.18) and (A.21), it can be seen that

$$Var(\boldsymbol{\varepsilon}'\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}\boldsymbol{\varepsilon}) \le 2tr(\mathbf{B}^{(st)}\mathbf{B}^{(st)\prime}) + C_1\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^{(st)*})\}^2 \le C'k_{S^*}$$

is valid uniformly in s and t, where C' is a constant unrelated to s and t. So, from Condition 6,

$$\begin{aligned} &\Pr\left\{\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon} - tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})|/R_n(\mathbf{w}) > \delta\right\} \\ &\le \sum_{s=1}^{S}\sum_{t=1}^{S}\Pr\{|\boldsymbol{\varepsilon}'\mathbf{P}^{(s)'}\mathbf{P}^{(t)}\boldsymbol{\varepsilon} - tr(\mathbf{P}^{(s)'}\mathbf{P}^{(t)}\boldsymbol{\Sigma})| > \delta\xi_n\} \\ &\le \delta^{-2}\xi_n^{-2}\sum_{s=1}^{S}\sum_{t=1}^{S}Var(\boldsymbol{\varepsilon}'\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}\boldsymbol{\varepsilon}) \\ &\le \delta^{-2}C'S^2k_{S^*}\xi_n^{-2} \\ &= \delta^{-2}C'(N^2k_{S^*}^2S^2n^{-2}\xi_n^{-2})N^{-2}n^2k_{S^*}^{-1} \to 0. \end{aligned}$$

Therefore, (A.16) is valid.

We now prove (A.17). From (1), we have $\|\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2 \le R_n(\mathbf{w})$. So

$$\begin{aligned}|\boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) &\le \left(\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon}\|\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2/R_n^2(\mathbf{w})\right)^{1/2} \\ &\le \left(\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon}/R_n(\mathbf{w})\right)^{1/2}. \end{aligned} \tag{A.22}$$

From the Cauchy–Schwarz inequality and Lemma 1(d), it is seen that

$$|tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)})| \le |tr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(s)})|^{1/2}|tr(\mathbf{P}^{(t)\prime}\mathbf{P}^{(t)})|^{1/2} \le C_3\sqrt{k_sk_t},$$

thus by Condition 5,

$$\begin{aligned} tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma}) &\le \bar{\sigma}(\boldsymbol{\Sigma})tr\left(\sum_{s=1}^{S}\sum_{t=1}^{S}w_sw_t\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}\right) \\ &= \bar{\sigma}(\boldsymbol{\Sigma})\sum_{s=1}^{S}\sum_{t=1}^{S}w_sw_ttr(\mathbf{P}^{(s)\prime}\mathbf{P}^{(t)}) \le C_2C_3k_{S^*}. \end{aligned} \tag{A.23}$$

Hence, from (A.16), (A.23) and Condition 6, we obtain

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon}|/R_n(\mathbf{w}) \\ &\le \sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\varepsilon} - tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})|/R_n(\mathbf{w}) \\ &\quad + \sup_{\mathbf{w}\in\mathcal{H}}|tr(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Sigma})|/R_n(\mathbf{w}) \xrightarrow{p} 0, \end{aligned}$$

which, along with (A.22), implies (A.17). The proof of Theorem 2.1 is completed.

Proof of Theorem 3.1. Let $\mathbf{M}^*(\mathbf{w}) = \mathbf{A}^{*\prime}(\mathbf{w})\mathbf{Q}^*(\mathbf{w}) + \mathbf{Q}^{*\prime}(\mathbf{w})\mathbf{A}^*(\mathbf{w}) + \mathbf{Q}^{*\prime}(\mathbf{w})\mathbf{Q}^*(\mathbf{w})$. Similar to the proof of Theorem 2.1, we need to verify

$$\sup_{\mathbf{w}\in\mathcal{H}}\bar{\sigma}(\mathbf{M}^*(\mathbf{w})) = O_p(1/T), \tag{A.24}$$

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}|/R_T^*(\mathbf{w}) \xrightarrow{p} 0, \tag{A.25}$$

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}|/R_T^*(\mathbf{w}) \xrightarrow{p} 0, \tag{A.26}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}|/R_T^*(\mathbf{w}) \xrightarrow{p} 0. \tag{A.27}$$

From the proof of Lemma 1(b), we see that $\bar{\sigma}(\mathbf{I} - \mathbf{P}^{(s)*}) \le 1$. Further, note that $\bar{\sigma}(\boldsymbol{\Pi}^{(s)}) = 1$ and $\bar{\sigma}(\boldsymbol{\Phi}^{(s)}) = \sqrt{2b_s+1}$. So following the steps of proving (A.1)–(A.4), (A.24) is valid.

Now we show that

$$\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)-1/2}\mathbf{X}^{(s)\prime}\boldsymbol{\varepsilon} \xrightarrow{d} N(0, \mathbf{I}). \tag{A.28}$$

Consider

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{X}_t^{(s)}\boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\zeta}$ is a real vector satisfying $\boldsymbol{\zeta}'\boldsymbol{\zeta} = 1$. It follows from Theorem 3.49 of White (1984) that $\boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{X}_t^{(s)}\boldsymbol{\varepsilon}_t$ is a sequence of α-mixing random variables with size $-\gamma/(\gamma-2)$ given Condition 8(i). According to Condition 8(ii) and (iii), we have

$$E(\boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{X}_t^{(s)}\varepsilon_t) = 0 \tag{A.29}$$

and

$$E|\boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{X}_t^{(s)}\varepsilon_t|^{\gamma} \le \Delta_1' < \infty, \tag{A.30}$$

where Δ_1' is a positive constant. From Condition 8(v), it is readily seen that

$$\begin{aligned} &Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{X}_t^{(s)}\boldsymbol{\varepsilon}_t\right) \\ &= \boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{G}_T^{(s)}\mathbf{G}_T^{(s)-1/2}\boldsymbol{\zeta} = 1. \end{aligned} \tag{A.31}$$

It follows from (A.29)–(A.31) and Theorem 5.20 of White (1984) that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\zeta}'\mathbf{G}_T^{(s)-1/2}\mathbf{X}_t^{(s)}\varepsilon_t \xrightarrow{d} N(0, 1). \tag{A.32}$$

Since (A.32) holds for any $\boldsymbol{\zeta}$ with $\boldsymbol{\zeta}'\boldsymbol{\zeta} = 1$, (A.28) is valid by virtue of Proposition 5.1 of White (1984).

Condition 8(i) implies that $\{\mathbf{X}_t^{(s)}\}$ is α-mixing of size $-\gamma/(\gamma-2)$. By Condition 8(iv)(a) and Corollary 3.48 of White (1984), we have $\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)}/T - \mathbf{H}_T^{(s)} \xrightarrow{a.s.} \mathbf{0}$. Given Condition 8(iv)(a), $\mathbf{H}_T^{(s)}$ is $O(1)$, which, together with Condition 8(iv)(b) and Proposition 2.30 of White (1984), indicates that $\left(\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)}/T\right)^{-1} - \mathbf{H}_T^{(s)^{-1}} \xrightarrow{p} \mathbf{0}$. Thus, we obtain

$$\bar{\sigma}\{(\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)}/T)^{-1}\} - \bar{\sigma}(\mathbf{H}_T^{(s)^{-1}}) \xrightarrow{p} 0. \tag{A.33}$$

Hence, from Condition 8(v), (A.28) and (A.33), it can be shown that

$$\begin{aligned}\boldsymbol{\varepsilon}'\mathbf{P}^{(s)^*}\boldsymbol{\varepsilon} &\le \bar{\sigma}\left\{\left(\frac{\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)} + \sum_{l=1}^{L}\lambda_l^{(s)}\mathbf{F}_l^{(s)}}{T}\right)^{-1}\right\} \\ &\quad\times \frac{1}{\sqrt{T}}\boldsymbol{\varepsilon}'\mathbf{X}^{(s)}\mathbf{G}_T^{(s)^{-1/2}}\mathbf{G}_T^{(s)}\mathbf{G}_T^{(s)^{-1/2}}\frac{1}{\sqrt{T}}\mathbf{X}^{(s)\prime}\boldsymbol{\varepsilon} \\ &\le \bar{\sigma}\left\{\left(\frac{\mathbf{X}^{(s)\prime}\mathbf{X}^{(s)}}{T}\right)^{-1}\right\}\bar{\sigma}(\mathbf{G}_T^{(s)})\left\|\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)^{-1/2}}\mathbf{X}^{(s)\prime}\boldsymbol{\varepsilon}\right\|^2 \\ &= O_p(1)\end{aligned} \tag{A.34}$$

uniformly in s. So by Condition 9 and the fact that S is bounded, we have

$$\sup_{\mathbf{w}\in\mathcal{H}}|\boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}|/R_T^*(\mathbf{w}) \le \xi_T^{*-1}\max_s|\boldsymbol{\varepsilon}'\mathbf{P}^{(s)^*}\boldsymbol{\varepsilon}| = o_p(1).$$

Therefore, (A.25) is valid.

Note that for each s, $\mathbf{P}^{(s)^*}$ is positive semi-definite and $\bar{\sigma}(\mathbf{P}^{(s)^*}) \le 1$. So $\mathbf{P}^*(\mathbf{w})$ is positive semi-definite and $\bar{\sigma}(\mathbf{P}^*(\mathbf{w})) \le 1$. Hence,

$$\boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon} \le \bar{\sigma}(\mathbf{P}^*(\mathbf{w}))\boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon} \le \boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}. \tag{A.35}$$

Thus, (A.26) is implied by (A.25).

Since

$$|\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}| \le \|\boldsymbol{\mu}\||\boldsymbol{\varepsilon}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\boldsymbol{\varepsilon}|^{1/2},$$

(A.27) holds by virtue of (A.34), (A.35) and Conditions 3 and 9. This completes the proof of Theorem 3.1.

Proof of Theorem 3.1 in the weakly correlated case where $E(\mathbf{X}_t^{(s)}\varepsilon_t) = O(\boldsymbol{\iota}_s/\sqrt{T})$

From the proof of Theorem 3.1, we see that if $\boldsymbol{\varepsilon}'\mathbf{P}^{(s)^*}\boldsymbol{\varepsilon} = O_p(1)$, then (8) is valid. For showing $\boldsymbol{\varepsilon}'\mathbf{P}^{(s)^*}\boldsymbol{\varepsilon} = O_p(1)$, following the steps of showing (A.28), we have

$$\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)^{-1/2}}(\mathbf{X}^{(s)\prime}\boldsymbol{\varepsilon} - E(\mathbf{X}^{(s)}\boldsymbol{\varepsilon})) \xrightarrow{d} N(0,\mathbf{I}). \tag{A.36}$$

Since $E(\mathbf{X}_t^{(s)}\varepsilon_t) = O(\boldsymbol{\iota}_s/\sqrt{T})$ uniformly for t,

$$\left\|\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)^{-1/2}}E(\mathbf{X}^{(s)}\boldsymbol{\varepsilon})\right\|^2 = O(1).$$

Consequently,

$$\begin{aligned}&\left\|\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)^{-1/2}}\mathbf{X}^{(s)\prime}\boldsymbol{\varepsilon}\right\|^2 \\ &\le \left\|\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)^{-1/2}}(\mathbf{X}^{(s)\prime}\boldsymbol{\varepsilon} - E(\mathbf{X}^{(s)}\boldsymbol{\varepsilon}))\right\|^2 + \left\|\frac{1}{\sqrt{T}}\mathbf{G}_T^{(s)^{-1/2}}E(\mathbf{X}^{(s)}\boldsymbol{\varepsilon})\right\|^2 \\ &= O_p(1),\end{aligned}$$

which, along with (A.34), implies $\boldsymbol{\varepsilon}'\mathbf{P}^{(s)^*}\boldsymbol{\varepsilon} = O_p(1)$.

References

Ando, T., Li, K.C., 2014. A model-averaging approach for high-dimensional regression. J. Amer. Statist. Assoc. 109, 254–265.
Antoniadis, A., 1997. Wavelets in statistics: A review. J. Ital. Stat. Assoc. 6, 97–144.
Arellano, M., 2003. Panel Data Econometrics. Oxford University Press, Oxford.
Baltagi, B., 2005. Econometric Analysis of Panel Data, second ed. Wiley, New York.
Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: An integral part of inference. Biometrics 53, 603–618.
Cheng, X., Hansen, B.E., 2015. Forecasting with factor-augmented regression: A frequentist model averaging approach. J. Econometrics 186, 280–293.
Claeskens, G., Croux, C., Venkerckhoven, J., 2006. Variable selection for logit regression using a prediction-focused information criterion. Biometrics 62, 972–979.
Claeskens, G., Krivobokova, T., Opsomer, J.D., 2009. Asymptotic properties of penalized spline estimators. Biometrika 96, 529–544.
Diggle, P.J., Heagerty, P.J., Liang, K.Y., Zeger, S.L., 2002. Analysis of Longitudinal Data, second ed. In: Oxford Statistical Science Series, vol. 25. Oxford University Press, Oxford.
Fan, J., 1997. Comments on 'wavelets in statistics: A review' by A. Antoniadis. J. Ital. Stat. Assoc. 6, 131–138.
Fan, J., Li, R., 2001. Variable selection via penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, J., Yao, Q., 2003. Nonlinear Time Series: Nonparametric and Parametric Methods. Springer-Verlag, New York.
Fan, J., Zhang, J.T., 2000. Two-step estimation of functional linear models with applications to longitudinal data. J. R. Stat. Soc. Ser. B Stat. Methodol 62, 303–322.
Green, P.J., Silverman, B.W., 1994. Nonparametric regression and generalized linear models: A roughness penalty approach. In: Monographs on Statistics and Applied Probability, vol. 58. Chapman & Hall, London.
Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.
Hansen, B.E., Racine, J.S., 2012. Jackknife model averaging. J. Econometrics 167, 38–46.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statist. Sci. 14, 382–417.
Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Amer. Statist. Assoc. 98, 879–899.
Hjort, N.L., Claeskens, G., 2006. Focussed information criteria and model averaging for Cox's hazard regression model. J. Amer. Statist. Assoc. 101, 1449–1464.
Hsiao, C., 2003. Analysis of Panel Data, second ed. Cambridge University Press, Cambriege.
Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.
Liang, H., Zou, G., Wan, A.T.K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. J. Amer. Statist. Assoc. 106, 1053–1066.
Lin, D.Y., Ying, Z., 2001. Semiparametric and nonparametric regression analysis of longitudinal data. J. Amer. Statist. Assoc. 96, 103–113.
Liu, C.-A., 2015. Distribution theory of the least squares averaging estimator. J. Econometrics 186, 142–159.
Rice, J.A., Silverman, B.W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. J. R. Stat. Soc. Ser. B Stat. Methodol 53, 233–243.
Ruppert, D., Wand, M., Carroll, R., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.
Tibshirani, R., 1996. Regression shrinkage and selection via lasso. J. R. Stat. Soc. Ser. B Stat. Methodol 58, 267–288.
Welsh, A.H., Lin, X., Carroll, R.J., 2002. Marginal longitudinal nonparametric regression: Locality and efficiency of spline and Kernel methods. J. Amer. Statist. Assoc. 97, 482–493.
White, H., 1984. Asymptotic Theory for Econometricians. Academic Press, Orlando.
Xu, G., Huang, J.Z., 2012. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. Ann. Statist. 40, 3003–3030.
Xu, G., Wang, S., Huang, J.Z., 2014. Focused information criterion and model averaging based on weighted composite quantile regression. Scand. J. Stat. 41, 365–381.
Yang, Y., 2001. Adaptive regression by mixing. J. Amer. Statist. Assoc. 96, 574–588.
Yuan, Z., Yang, Y., 2005. Combining linear regression models: When and how? J. Amer. Statist. Assoc. 100, 1202–1214.
Zeger, S.L., Diggle, P.J., 1994. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics 50, 689–699.
Zhang, D., Lin, X., Raz, J., Sowers, M., 1998. Semiparametric stochastic mixed models for longitudinal data. J. Amer. Statist. Assoc. 93, 710–719.
Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Ann. Statist. 39, 174–200.
Zhang, X., Wan, A.T.K., Zhou, S.Z., 2012. Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. J. Bus. Econom. Statist. 30, 132–142.
Zhang, X., Wan, A.T.K., Zou, G., 2013. Model averaging by Jackknife criterion in models with dependent data. J. Econometrics 174, 82–94.
Zhang, X., Zou, G., Liang, H., 2014. Model averaging and weight choice in linear mixed-effects models. Biometrika 101, 205–218.
Zhu, Z., Fung, W.K., He, X., 2008. On the asymptotics of marginal regression splines with longitudinal data. Biometrika 95, 907–917.