Submitted to the Annals of Statistics
arXiv: math.PR/0000000

ASYMPTOTIC OPTIMALITY AND EFFICIENT COMPUTATION OF THE LEAVE-SUBJECT-OUT CROSS-VALIDATION

BY GANGGANG XU AND JIANHUA Z. HUANG*

Texas A&M University

Although the leave-subject-out cross-validation (CV) has been widely used in practice for tuning parameter selection for various nonparametric and semiparametric models of longitudinal data, its theoretical property is unknown and solving the associated optimization problem is computationally expensive, especially when there are multiple tuning parameters. In this paper, by focusing on the penalized spline method, we show that the leave-subject-out CV is optimal in the sense that it is asymptotically equivalent to the empirical squared error loss function minimization. An efficient Newton-type algorithm is developed to compute the penalty parameters that optimize the CV criterion. Simulated and real data are used to demonstrate the effectiveness of the leave-subject-out CV in selecting both the penalty parameters and the working correlation matrix.

1. Introduction. In recent years there have seen growing interests in applying flexible statistical models for analyzing longitudinal data or the more general cluster data. Various semiparametric (e.g., Zeger and Diggle, 1994; Zhang et al., 1998; Lin and Ying, 2001; Wang et al., 2005) and nonparametric (e.g., Rice and Silverman, 1991; Wang, 1998; Fan and Zhang, 2000; Lin and Carroll, 2000; Welsh et al., 2002; Wang, 2003; Zhu et al., 2008) models have been proposed and studied in the literature. All of these flexible, semiparametric or nonparametric methods require specification of tuning parameters, such as the bandwidth for the local polynomial kernel methods, the number of knots for regression splines, and the penalty parameter for penalized splines and smoothing splines.

The "leave-subject-out cross-validation" (LsoCV) or more generally called "leave-cluster-out cross-validation", introduced by Rice and Silverman (1991), has been widely used as the method for selecting tuning parameters in analyzing longitudinal data and clustered data; see, for example, Hoover et al.

*Corresponding author. Research was partly supported by NSF (DMS-0907170, DMS-1007618, DMS-1208952), NCI (CA57030), and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

Keywords and phrases: cross-validation, generalized estimating equations, multiple smoothing parameters, penalized splines, working correlation matrix

(1998); Huang et al. (2002); Wu and Zhang (2006); Wang et al. (2008). The LsoCV is intuitively appealing since the within-subject dependence is preserved by leaving out all observations from the same subject together in the cross-validation. In spite of its broad acceptance in practice, the use of LsoCV still lacks a theoretical justification to date. Computationally, the existing literature has focused on the grid search method for finding the minimizer of the LsoCV criterion (LsoCV score) (Chiang et al., 2001; Huang et al., 2002; Wang et al., 2008), which is rather inefficient and even prohibitive with the existence of multiple tuning parameters. The goal of this paper is twofold: First, we develop a theoretical justification of the LsoCV by showing that the LsoCV criterion is asymptotically equivalent to an appropriately defined loss function; second, we develop a computationally efficient algorithm to optimize the LsoCV criterion for selecting multiple penalty parameters for penalized splines.

We shall focus our presentation on longitudinal data, but all discussions in this paper apply to clustered data analysis. Suppose we have n subjects and each subject has a series of observations $(y_{ij}, \boldsymbol{x}_{ij})$, for $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, with y_{ij} being the jth response from the ith subject and $\boldsymbol{x}_{ij}$ being the corresponding vector of covariates. Denote $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^T$ and $\tilde{\mathbf{X}}_i = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{in_i})$. The marginal non- and semi-parametric regression model (Welsh et al., 2002; Zhu et al., 2008) assumes that the mean and covariance matrix of the responses are given by

$$\mu_{ij} = E(y_{ij}|\tilde{\mathbf{X}}_i) = \boldsymbol{x}_{ij0}\boldsymbol{\beta}_0 + \sum_{k=1}^{m} f_k(\boldsymbol{x}_{ijk}), \quad cov(\boldsymbol{y}_i|\tilde{\mathbf{X}}_i) = \boldsymbol{\Sigma}_i, \tag{1}$$

where $\boldsymbol{\beta}_0$ is a vector of linear regression coefficients, f_k, $k = 1, \ldots, m$, are unknown smooth functions, and $\boldsymbol{\Sigma}_i$'s are within-subject covariance matrices. Denote $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{in_i})^T$. By using a basis expansion to approximate each f_k, $\boldsymbol{\mu}_i$ can be approximated by $\boldsymbol{\mu}_i \approx \mathbf{X}_i\boldsymbol{\beta}$ for some matrix $\mathbf{X}_i$ and unknown parameter vector $\boldsymbol{\beta}$, which then can be estimated by minimizing the penalized weighted least square loss function

$$pl(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{W}_i^{-1} (\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{k=1}^{m} \lambda_k \boldsymbol{\beta}^T \mathbf{S}_k \boldsymbol{\beta}, \tag{2}$$

where $\mathbf{W}_i$'s are working correlation matrices that are possibly misspecified, $\mathbf{S}_k$ is a semi-positive definite matrix such that $\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta}$ serves as a roughness penalty for f_k, and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)$ is a vector of penalty parameters.

Methods for choosing basis functions, constructing the corresponding design matrices $\mathbf{X}_i$, and defining the roughness penalty matrices are well established in the statistics literature. For example, B-spline basis and basis obtained from reproducing kernel Hilbert spaces are commonly used. Roughness penalty matrices can be formed corresponding to the squared second-difference penalty, the squared second derivative penalty, the thin-plate splines penalty, or using directly the reproducing kernels. We refer to the books by Green and Silverman (1994), Gu (2002), and Wood (2006) for thorough treatments of this subject.

The idea of using working correlation for longitudinal data can be traced back to the generalized estimating equations (GEE) of Liang and Zeger (1986), where it is established that the mean function can be consistently estimated with the correct inference even when the correlation structure is misspecified. Liang and Zeger (1986) further demonstrated that using a possibly misspecified working correlation structure $\mathbf{W}$ has the potential to improve the estimation efficiency over methods that completely ignore the within-subject correlation. Similarly results have been obtained in the nonparametric setting in Welsh et al. (2002) and Zhu et al. (2008). Commonly used working correlation structures include compound symmetry, autoregressive models; see Diggle et al. (2002) for a detailed discussion.

In the case of independent data, Li (1986) established the asymptotic optimality of the generalized cross validation (GCV) (Craven and Wahba, 1979) for penalty parameter selection by showing that minimizing the GCV criterion is asymptotically equivalent to minimizing a suitably defined loss function. To understand the theoretical property of LsoCV, we ask the following question in this paper: What loss function does the LsoCV mimic or estimate and how good is this estimation? We are able to show that the unweighted mean squared error is the loss function that LsoCV is targeting. Specifically, we obtain that, up to a quantity that does not depend on the penalty parameters, the LsoCV score is asymptotically equivalent to the mean squared error loss. Our result provides the needed theoretical justification of the wide use of LsoCV in practice.

In two related papers, Gu and Ma (2005) and Han and Gu (2008) developed modifications of GCV for dependent data under assumptions on the correlation structure and established the optimality of the modified GCV. Although their modified GCVs work well when the correlation structure is correctly specified up to some parameters, they need not be suitable when there is not enough prior knowledge to make such a specification or the within-subject correlation is too complicated to be modeled nicely with a simple structure. The main difference between LsoCV and these modified GCVs is that LsoCV utilizes working correlation matrices in the estimating equations and allows mis-specification of the correlation structure. Moreover, since the LsoCV and the asymptotic equivalent squared error loss are not attached to any specific correlation structure, LsoCV can be used to select not only the penalty parameters but also the correlation structure.

Another contribution of this paper is the development of a fast algorithm for optimizing the LsoCV criterion. To avoid computation of a large number of matrix inversions, we first derive an asymptotically equivalent approximation of the LsoCV criterion and then derive a Newton–Raphson type algorithm to optimize this approximated criterion. The algorithm is particularly useful when we need to select multiple penalty parameters.

The rest of the paper is organized as follows. Section 2 presents the main theoretical results. Section 3 proposes a computationally efficient algorithm for optimizing the LosCV criterion. Results from some simulation studies and a real data analysis are given in Sections 4 and 5. All technical proofs and computational implementations are collected in the Appendix and in the Supplementary materials.

2. Leave-subject-out Cross Validation. Let $\hat{\mu}(\cdot)$ denote the estimate of the mean function obtained by using basis expansion of unknown functions f_k's ($k = 1, \dots, m$) and solving the minimization problem (2) for $\boldsymbol{\beta}$. Let $\hat{\mu}^{[-i]}(\cdot)$ be the estimate of the mean function $\mu(\cdot)$ by the same method but using all the data except observations from subject i, $1 \le i \le n$. The LsoCV criterion is defined as

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{y}_i - \hat{\mu}^{[-i]}(\mathbf{X}_i)\}^T\{\boldsymbol{y}_i - \hat{\mu}^{[-i]}(\mathbf{X}_i)\}. \tag{3}$$

By leaving out together all observations from the same subject, the within-subject correlation is preserved in LsoCV. Before giving the formal justification of LsoCV, we review a heuristic justification in Section 2.1. Section 2.2 defines the suitable loss function. Section 2.3 lists the regularity conditions and Section 2.4 provides an example illustrating how the regularity conditions in Section 2.3 can be verified using more primitive conditions. Section 2.5 presents the main theoretical result about the optimality of LsoCV.

2.1. *Heuristic justification.* The initial heuristic justification of LsoCV by Rice and Silverman (1991) is that it mimics the mean squared prediction error (MSPE). Consider some new observations $(\mathbf{X}_i, \boldsymbol{y}_i^*)$, taken at the same design points as the observed data. For a given estimator of the mean function $\mu(\cdot)$, denoted as $\hat{\mu}(\cdot)$, the MSPE is defined as

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n} E\|\boldsymbol{y}_i^* - \hat{\mu}(\mathbf{X}_i)\|^2 = \frac{1}{n}tr(\boldsymbol{\Sigma}) + \frac{1}{n}\sum_{i=1}^{n} E\|\mu(\mathbf{X}_i) - \hat{\mu}(\mathbf{X}_i)\|^2.$$

Using the independence between $\hat{\mu}^{[-i]}(\cdot)$ and $\boldsymbol{y}_i$ we obtain that

$$E\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})\} = \frac{1}{n}tr(\boldsymbol{\Sigma}) + \frac{1}{n}\sum_{i=1}^{n} E\|\mu(\mathbf{X}_i) - \hat{\mu}^{[-i]}(\mathbf{X}_i)\|^2,$$

where $\boldsymbol{\Sigma} = diag\{\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_n\}$. When n is large, $\hat{\mu}^{[-i]}(\cdot)$ should be close to $\hat{\mu}(\cdot)$, the estimate that uses observations from all subjects. Thus, we would expect that $E\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})\}$ to be close to the MSPE.

2.2. *Loss function.* We shall provide a formal justification of LsoCV by showing that the LsoCV is asymptotically equivalent to an appropriately defined loss function. Denote $\boldsymbol{Y} = (\boldsymbol{y}_1^T,\cdots,\boldsymbol{y}_n^T)^T$, $\mathbf{X} = (\mathbf{X}_1^T,\cdots,\mathbf{X}_n^T)^T$, and $\mathbf{W} = diag\{\mathbf{W}_1,\cdots,\mathbf{W}_n\}$. Then, for a given choice of $\boldsymbol{\lambda}$ and $\mathbf{W}$, the minimizer of (2) has a closed-form expression

$$\hat{\boldsymbol{\beta}} = \Big(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X} + \sum_{k=1}^{m}\lambda_k\mathbf{S}_k\Big)^{-1}\mathbf{X}^T\mathbf{W}^{-1}\boldsymbol{Y}. \tag{4}$$

The fitted mean function evaluated at the design points is given by

$$\hat{\mu}(\mathbf{X}|\boldsymbol{Y},\mathbf{W},\boldsymbol{\lambda}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{A}(\mathbf{W},\boldsymbol{\lambda})\boldsymbol{Y}, \tag{5}$$

where $\mathbf{A}(\mathbf{W},\boldsymbol{\lambda})$ is the hat matrix defined as

$$\mathbf{A}(\mathbf{W},\boldsymbol{\lambda}) = \mathbf{X}\Big(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X} + \sum_{k=1}^{m}\lambda_k\mathbf{S}_k\Big)^{-1}\mathbf{X}^T\mathbf{W}^{-1}. \tag{6}$$

From now on, we shall use $\mathbf{A}$ for $\mathbf{A}(\mathbf{W},\boldsymbol{\lambda})$ without causing any confusion.

For a given estimator $\hat{\mu}(\cdot)$ of $\mu(\cdot)$, define the mean squared error (MSE) loss as the true loss function

$$L(\hat{\boldsymbol{\mu}}) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(\mathbf{X}_i) - \mu(\mathbf{X}_i)\}^T\{\hat{\mu}(\mathbf{X}_i) - \mu(\mathbf{X}_i)\}. \tag{7}$$

Using (5), we obtain that, for the estimator obtained by minimizing (2), the true loss function (7) becomes

$$\begin{aligned} L(\mathbf{W},\boldsymbol{\lambda}) &= \frac{1}{n}(\mathbf{A}\boldsymbol{Y}-\boldsymbol{\mu})^T(\mathbf{A}\boldsymbol{Y}-\boldsymbol{\mu}) \\ &= \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon} - \frac{2}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A}^T)\mathbf{A}\boldsymbol{\epsilon}, \end{aligned} \tag{8}$$

where $\boldsymbol{\mu} = (\mu(\mathbf{X}_1)^T,\cdots,\mu(\mathbf{X}_n)^T)^T$, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$. Since $E(\boldsymbol{\epsilon}|\tilde{\mathbf{X}}_1,\ldots,\tilde{\mathbf{X}}_n) = 0$ and $Var(\boldsymbol{\epsilon}|\tilde{\mathbf{X}}_1,\ldots,\tilde{\mathbf{X}}_n) = \boldsymbol{\Sigma}$, the risk function can be derived as

$$R(\mathbf{W},\boldsymbol{\lambda}) = E\{L(\mathbf{W},\boldsymbol{\lambda})\} = \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}). \tag{9}$$

2.3. *Regularity conditions.* This section states some regularity conditions needed for our theoretical results. Noticing that unless $\mathbf{W} = \mathbf{I}$, $\mathbf{A}$ is not symmetric. We define a symmetric version of $\mathbf{A}$ as $\tilde{\mathbf{A}} = \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{1/2}$. Let $\mathbf{C}_{ii}$ be the diagonal block of $\tilde{\mathbf{A}}^2$ corresponding to the ith subject. With some abuse of notations (but clear from the context), denote by $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ the largest and the smallest eigenvalues of a matrix. The regularity conditions invovle the quantity $\xi(\mathbf{\Sigma}, \mathbf{W}) = \lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})$, which takes the minimal value $\lambda_{max}(\mathbf{\Sigma})$ when $\mathbf{W} = \mathbf{I}$ or $\mathbf{W} = \mathbf{\Sigma}$. Let $\boldsymbol{e}_i = \mathbf{\Sigma}_i^{-1/2}\boldsymbol{\epsilon}_i$ and $\boldsymbol{u}_i$ be $n_i \times 1$ vectors such that $\boldsymbol{u}_i^T\boldsymbol{u}_i = 1$, $i = 1, \cdots, n$.

Condition 1. For some $K > 0$, $E\{(\boldsymbol{u}_i^T\boldsymbol{e}_i)^4\} \leq K$, $i = 1, \ldots, n$.
Condition 2.

(i) $\max_{1\leq i\leq n}\{tr(\mathbf{A}_{ii})\} = O(tr(\mathbf{A})/n) = o(1)$;
(ii) $\max_{1\leq i\leq n}\{tr(\mathbf{C}_{ii})\} = o(1)$.

Condition 3. $\xi(\mathbf{\Sigma}, \mathbf{W})/n = o(R(\mathbf{W}, \boldsymbol{\lambda}))$.
Condition 4. $\xi(\mathbf{\Sigma}, \mathbf{W})\{n^{-1}tr(\mathbf{A})\}^2/\{n^{-1}tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma})\} = o(1)$.
Condition 5. $\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2) = o(1)$.

Condition 1 is a mild moment condition that requires that each component of the standardized residual $\boldsymbol{e}_i = \Sigma_i^{-1/2}\boldsymbol{\epsilon}_i$ has uniformly bounded fourth moment. In particular, when $\boldsymbol{\epsilon}_i$'s are from the Gaussian distribution, the condition holds with $K = 3$.

Condition 2 extends the usual condition on leverage points used in theoretical analysis of linear regression models. Note that $\{tr(\mathbf{A}_{ii})\}$ can be interpreted as the leverage of subject i, measuring the contribution to the fit from data of subject i and the average of the leverages is $tr(\mathbf{A})/n$. This condition says that the maximum leverage can not be arbitrarily larger than the average leverage, or in other words, there should not be any dominant or extremely influential subjects. In the special case that all subjects have the same design matrices, the condition automatically satisfies since $tr(\mathbf{A}_{ii}) = tr(\mathbf{A})/n$ for all $i = 1, \cdots, n$. Condition 2 is likely to be violated if the n_i's are very unbalanced. For example, if 10% of subjects have 20 observations and the rest of subjects only has 2 or 3 observations each, then $\max_{1\leq i\leq n}\{tr(\mathbf{A}_{ii})\}/\{n^{-1}tr(\mathbf{A})\}$ can be very large.

When n_i's are bounded, any reasonable choice of $\mathbf{W}$ would generally yield a bounded value of the quantity $\xi(\mathbf{\Sigma}, \mathbf{W})$, and Condition 3 reduces to $nR(\mathbf{W}, \boldsymbol{\lambda}) \to \infty$, which simply says that the parametric rate of $O(n^{-1})$ is not achievable. This is a mild condition since we are considering nonparametric estimation. When n_i's are not bounded, Condition 3's verification should be done on a case-by-case basis. As a special case, recent results for the longitudinal function estimation by Cai and Yuan (2011) indicate that Condition 3 would be satisfied in this particular setting if $\xi(\mathbf{\Sigma}, \mathbf{W})/n^* = O(1)$ and $n^*/n^{1/2r} \to 0$ or $\xi(\mathbf{\Sigma}, \mathbf{W})/n^* = o(1)$ and $n^*/n^{1/2r} \to \infty$ for some $r > 1$, where $n^* = (\frac{1}{n}\sum_{i=1}^n \frac{1}{n_i})^{-1}$ is the harmonic mean of $n_1, \ldots, n_n$. This conclusion holds for both fixed common designs and independent random designs.

Condition 4 essentially says that $\xi(\mathbf{\Sigma}, \mathbf{W})\{n^{-1}tr(\mathbf{A})\}^2 = o(R(\mathbf{W}, \boldsymbol{\lambda}))$. It is straightforward to show that the left-hand side is bounded from above by $c(\mathbf{\Sigma}\mathbf{W}^{-1})c(\mathbf{W})\{tr(\tilde{\mathbf{A}})/n\}^2/\{tr(\tilde{\mathbf{A}}^2)/n\}$, where $c(\mathbf{M}) = \lambda_{max}(\mathbf{M})/\lambda_{min}(\mathbf{M})$ is the condition number of a matrix $\mathbf{M}$. If n_i's are bounded, for choices of $\mathbf{W}$ such that $\mathbf{\Sigma}\mathbf{W}^{-1}$ and $\mathbf{W}$ are not singular, it suffices to guarantee $\{tr(\tilde{\mathbf{A}})/n\}^2/\{tr(\tilde{\mathbf{A}}^2)/n\} = o(1)$. For regression splines ($\boldsymbol{\lambda} = \mathbf{0}$), this condition holds if $p/n \to 0$ where p is the number of basis functions used, since $tr(\tilde{\mathbf{A}}^2) = tr(\tilde{\mathbf{A}}) = p$. For penalized splines and smoothing splines, we provide a more detailed discussion in Section 2.4.

If the working correlation matrix $\mathbf{W}$ is chosen to be well-conditioned such that its condition number $\lambda_{max}(\mathbf{W})/\lambda_{min}(\mathbf{W})$ is bounded, Condition 5 reduces to $tr(\mathbf{A})/n \to 0$, which can be verified as Condition 4.

Conditions 3–5 all indicate that a bad choice of the working correlation matrix $\mathbf{W}$ may deteriorate the performance of the LsoCV method. For example, Conditions 3–5 may be violated when $\mathbf{\Sigma}^{-1}\mathbf{W}$ or $\mathbf{W}$ is nearly singular. Thus in practice, it is wise to avoid using working correlation $\mathbf{W}$ that is nearly singular.

We do not make the assumption that n_i's are bounded. However, n_i obviously can not grow too fast relative to the number of subjects n. In particular, if n_i's are too large, $\lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1})$ can be fairly large unless $\mathbf{W} \approx \mathbf{\Sigma}$, and $\lambda_{max}(\mathbf{W})$ can be fairly large due to increase of dimensions of the working correlation matrices for individual subjects. Thus, Conditions 3–5 implicitly impose a limit to the growth rate of n_i.

2.4. *An example: penalized splines with B-spline basis functions.* In this section, we provide an example where Conditions 3–5 can be discussed in a more specific manner. Consider Model (1) with only one nonparametric covariate x and thus only one penalty parameter λ. We further assume that all eigeinvalues of matrices $\mathbf{W}$ and $\mathbf{\Sigma}\mathbf{W}^{-1}$ are bounded from below and above, that is, there exist positive constants c_1 and c_2 such that $c_1 \le \lambda_{min}(\mathbf{W}) \le \lambda_{max}(\mathbf{W}) \le c_2$ and $c_1 \le \lambda_{min}(\mathbf{\Sigma}\mathbf{W}^{-1}) \le \lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1}) \le c_2$. Under this assumption, it is straightforward to show that Conditions 3–5 reduce to the following conditions.

Condition $3'$. $nR(\mathbf{W}, \lambda) \to \infty$ as $n \to \infty$.
Condition $4'$. $\{n^{-1}tr(\mathbf{A})\}^2/\{n^{-1}tr(\tilde{\mathbf{A}}^2)\} = o(1)$.

Condition $5'$. $tr(\mathbf{A})/n = o(1)$.

Using Lemmas 4.1 and 4.2 from Han and Gu (2008) and similar arguments, we have the following three inequalities

$$tr\{\tilde{\mathbf{A}}(c_2\lambda, \mathbf{I})\} \le tr\{\tilde{\mathbf{A}}(\lambda, \mathbf{W})\} \le tr\{\tilde{\mathbf{A}}(c_1\lambda, \mathbf{I})\}, \tag{10}$$

$$tr\{\tilde{\mathbf{A}}^2(c_2\lambda, \mathbf{I})\} \le tr\{\tilde{\mathbf{A}}^2(\lambda, \mathbf{W})\} \le tr\{\tilde{\mathbf{A}}^2(c_1\lambda, \mathbf{I})\}, \tag{11}$$

and

$$\begin{aligned} &c_1 c_3^{-1}\{\mathbf{I} - \tilde{\mathbf{A}}(c_2\lambda, \mathbf{I})\} \\ &\qquad \le \{\mathbf{I} - \mathbf{A}(\lambda, \mathbf{W})\}^T\{\mathbf{I} - \mathbf{A}(\lambda, \mathbf{W})\} \le c_2 c_3\{\mathbf{I} - \tilde{\mathbf{A}}(c_2\lambda, \mathbf{I})\}, \end{aligned} \tag{12}$$

where $c_3 = \exp\{c_2(1 + (c_1^{-1} - c_2^{-1})^2 + (c_1^{-1} - c_2^{-1}))\}$. These inequalities and the definition of the risk function $R(\mathbf{W}, \lambda)$ imply that we need only to check Conditions $3'$–$5'$ for the case that $\mathbf{W} = \mathbf{I}$. In particular, (10)–(12) imply that

$$\begin{aligned} &c_1 c_3^{-1}\boldsymbol{\mu}^T\{\mathbf{I} - \tilde{\mathbf{A}}(c_2\lambda, \mathbf{I})\}^2\boldsymbol{\mu} + c_1^2 tr\{\tilde{\mathbf{A}}^2(c_2\lambda, \mathbf{I})\} \\ &\qquad \le nR(\mathbf{W}, \lambda) \le c_2 c_3\boldsymbol{\mu}^T\{\mathbf{I} - \tilde{\mathbf{A}}(c_1\lambda, \mathbf{I})\}^2\boldsymbol{\mu} + c_2^2 tr\{\tilde{\mathbf{A}}^2(c_1\lambda, \mathbf{I})\}, \end{aligned}$$

and therefore to show Condition $3'$, it suffices to show

$$\boldsymbol{\mu}^T\{\mathbf{I} - \tilde{\mathbf{A}}(\lambda, \mathbf{I})\}^2\boldsymbol{\mu} \to \infty \quad \text{or} \quad tr\{\tilde{\mathbf{A}}^2(\lambda, \mathbf{I})\} \to \infty \tag{13}$$

as $n \to \infty$.

We now use existing results from the literature to show how to verify Conditions $3'$–$5'$. Note that the notations used in the literature of penalized splines and smoothing splines are not always consistent. To fix notation, we denote for the rest of this section that $\lambda^* = \lambda/N$ and $\tilde{\mathbf{A}}^*(\lambda^*) = \tilde{\mathbf{A}}(\lambda, \mathbf{I})$, where N is the total number of observations from all subjects.

Let r denote the order of the B-splines and consider a sequence of knots defined on the interval $[a, b]$, $a = t_{-(r-1)} = \cdots = t_0 < t_1 < \cdots < t_{K_n} < t_{K_n+1} = \cdots = t_{K_n+r} = b$. Define B-spline basis functions recursively as

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

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

for $j = -(r-1), \ldots, K_n$. When this B-spline basis is used for basis expansion, the jth row of $\mathbf{X}_i$ is $\mathbf{X}_{i(j)}^T = (B_{-(r-1),r}(x_{ij}), \ldots, B_{K_n,r}(x_{ij}))$, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. When the penalty is the integrated squared qth derivative of the spline function with $q \le r-1$, i.e., $\int (f^{(q)})^2$, the penalty term can be written in terms of the spline coefficient vector $\boldsymbol{\beta}$ as $\lambda \boldsymbol{\beta}^T \Delta_q^T R \Delta_q \boldsymbol{\beta}$, where R is a $(K_n+r-q) \times (K_n+r-q)$ matrix with $R_{ij} = \int_a^b B_{j,r-q}(x) B_{i,r-q}(x) dx$ and Δ_q is a matrix of weighted qth order difference operator (Claeskens et al., 2009).

We make the following assumptions: (a) $\delta = \max_{0 \le j \le K_n}(t_{j+1} - t_j)$ is of the order $O(K_n^{-1})$ and $\delta / \min_{0 \le j \le K_n}(t_{j+1} - t_j) \le M$ for some constant $M > 0$; (b) $\sup_{x \in [a,b]} |Q_n(x) - Q(x)| = o(K_n^{-1})$ where Q_n and Q are empirical and true distribution function of all design points $\{x_1, \ldots, x_N\}$; (c) $K_n = o(N)$. Define quantity $K_q = (K_n+r-q)(\lambda^* \tilde{c}_1)^{1/(2q)}$ with some constant $\tilde{c}_1 > 0$ depending on q and the design density. Claeskens et al. (2009) showed that, under above assumptions, if $K_q < 1$, $tr\{\tilde{\mathbf{A}}^*(\lambda^*)\}$ and $tr\{\tilde{\mathbf{A}}^{*2}(\lambda^*)\}$ are both of the order $O(K_n)$ and $\boldsymbol{\mu}^T \{\mathbf{I} - \tilde{\mathbf{A}}^*(\lambda^*)\}^2 \boldsymbol{\mu} = O(\lambda^{*2} N K_n^{2q} + N K_n^{-2r})$; If $K_q \ge 1$, $tr\{\tilde{\mathbf{A}}^*(\lambda^*)\}$ and $tr\{\tilde{\mathbf{A}}^{*2}(\lambda^*)\}$ are of order $O(\lambda^{*-1/(2q)})$ and $\boldsymbol{\mu}^T \{\mathbf{I} - \tilde{\mathbf{A}}^*(\lambda^*)\}^2 \boldsymbol{\mu} = O(N\lambda^* + N K_n^{-2q})$. Using these results and the results following inequalities (10)–(12), it is straightforward to show that if $\lambda^* = 0$ (for regression splines), letting $K_n \to \infty$ and $K_n/n \to 0$ is sufficient to guarantee Conditions $3'$–$5'$, and if $\lambda^* \neq 0$ (for penalized splines), further assuming $\lambda^* \to 0$ and $n\lambda^{*1/(2q)} \to \infty$ ensures the validity of Conditions $3'$–$5'$.

It is noticeable that when $K_q \ge 1$, the asymptotic property of the penalized spline estimator is close to that of smoothing splines, where the number of internal knots $K_n = N$. In fact, as discussed in Han and Gu (2008), for smoothing splines, it typically holds that $tr\{\tilde{\mathbf{A}}^*(\lambda^*)\}$ and $tr\{\tilde{\mathbf{A}}^{*2}(\lambda^*)\}$ are of order $O(\lambda^{*-1/d})$ and $\boldsymbol{\mu}^T \{\mathbf{I} - \tilde{\mathbf{A}}^*(\lambda^*)\}^2 \boldsymbol{\mu} = O(N\lambda^*)$ for some $d > 1$ as $N \to \infty$ and $\lambda^* \to 0$; see also Craven and Wahba (1979), Li (1986), Gu (2002). Therefore, if one has $\lambda^* \to 0$ and $n\lambda^{*1/d} \to \infty$, Conditions $3'$–$5'$ can be verified for smoothing splines.

2.5. *Optimality of Leave-subject-out CV.* In this subsection, we provide a theoretical justification of using the minimizer of $\mathrm{LosCV}(\mathbf{W}, \boldsymbol{\lambda})$ to select the optimal value of the penalty parameters $\boldsymbol{\lambda}$. We say that the working correlation matrix $\mathbf{W}$ is pre-determined if it is determined by observation times and/or some other covariates. One way to obtain such $\mathbf{W}$ is to use some correlation function plugged in with estimated parameters. Naturally, it is reasonable to consider the value of $\boldsymbol{\lambda}$ that minimizes the true loss function $L(\mathbf{W}, \boldsymbol{\lambda})$ as the optimal value of the penalty parameters for a pre-determined $\mathbf{W}$. However, $L(\mathbf{W}, \boldsymbol{\lambda})$ can not be evaluated using data alone since the true mean function in the definition of $L(\mathbf{W}, \boldsymbol{\lambda})$ is unknown. One idea is to use an unbiased estimate of the risk function $R(\mathbf{W}, \boldsymbol{\lambda})$ as a proxy of $L(\mathbf{W}, \boldsymbol{\lambda})$. Define

$$U(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I} - \mathbf{A})^T(\mathbf{I} - \mathbf{A})\mathbf{Y} + \frac{2}{n}tr(\mathbf{A}\boldsymbol{\Sigma}). \tag{14}$$

It is easy to show that

$$U(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \frac{2}{n}\boldsymbol{\mu}^T(\mathbf{I} - \mathbf{A})^T\boldsymbol{\epsilon} - \frac{2}{n}\{\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon} - tr(\mathbf{A}\boldsymbol{\Sigma})\}, \tag{15}$$

which has expectation zero. Thus, if $\boldsymbol{\Sigma}$ is known, $U(\mathbf{W}, \boldsymbol{\lambda}) - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ is an unbiased estimate of the risk $R(\mathbf{W}, \boldsymbol{\lambda})$. Actually, the estimator is consistent, as stated in the following theorem.

THEOREM 2.1. *Under Conditions* 1–4*, for a pre-determined* $\mathbf{W}$ *and a non-random* $\boldsymbol{\lambda}$*, as* $n \to \infty$*,*

$$L(\mathbf{W}, \boldsymbol{\lambda}) - R(\mathbf{W}, \boldsymbol{\lambda}) = o_p(R(\mathbf{W}, \boldsymbol{\lambda}))$$

and

$$U(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W}, \boldsymbol{\lambda})).$$

This theorem shows that, the function $U(\mathbf{W}, \boldsymbol{\lambda}) - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$, the loss function $L(\mathbf{W}, \boldsymbol{\lambda})$, and the risk function $R(\mathbf{W}, \boldsymbol{\lambda})$ are asymptotically equivalent. Thus, if $\boldsymbol{\Sigma}$ is known, $U(\mathbf{W}, \boldsymbol{\lambda}) - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ is a consistent estimator of the risk function and moreover, $U(\mathbf{W}, \boldsymbol{\lambda})$ can be used as a reasonable surrogate of $L(\mathbf{W}, \boldsymbol{\lambda})$ for selecting the penalty parameters, since the $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ term does not depend on $\boldsymbol{\lambda}$. However, $U(\mathbf{W}, \boldsymbol{\lambda})$ depends on knowledge of the true covariance matrix $\boldsymbol{\Sigma}$, which is usually not available. The following result states that the LsoCV score provides a good approximation of $U(\mathbf{W}, \boldsymbol{\lambda})$, without the knowledge of $\boldsymbol{\Sigma}$.

THEOREM 2.2. *Under Conditions* 1–5*, for a pre-determined* $\mathbf{W}$ *and a non-random* $\boldsymbol{\lambda}$*, as* $n \to \infty$*,*

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) - U(\mathbf{W}, \boldsymbol{\lambda}) = o_p(L(\mathbf{W}, \boldsymbol{\lambda})),$$

and therefore

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W}, \boldsymbol{\lambda})).$$

This theorem suggests that minimizing $\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is asymptotically equivalent to minimizing $U(\mathbf{W}, \boldsymbol{\lambda})$ and is also equivalent to minimizing the true loss function $L(\mathbf{W}, \boldsymbol{\lambda})$. Unlike $U(\mathbf{W}, \boldsymbol{\lambda})$, $\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})$ can be evaluated using the data. The theorem provides the justification of using LsoCV, as a consistent estimator of the loss or risk function, for selecting the penalty parameters.

Remark 1. Although the results are presented for selection of the penalty parameter $\boldsymbol{\lambda}$ for penalized splines, the results also hold for selection of knot numbers (or number of basis functions) K_n for regression splines when $\boldsymbol{\lambda} = \mathbf{0}$ and K_n is the tuning parameter to be selected.

Remark 2. Since the definition of the true loss function (7) does not depend on the working correlation structure $\mathbf{W}$, we can use this loss function to compare performances of different choices of $\mathbf{W}$, for example, compound symmetry or autoregressive, and then choose the best one among several candidates. Thus, the result in Theorem 2.2 also suggests and provides a justification to use the LsoCV for selecting the working correlation matrix. This suggestion is evaluated using a simulation study in Section 4.3. When using the LsoCV to select the working correlation matrix, we recommend to use regression splines, i.e. setting $\boldsymbol{\lambda} = \mathbf{0}$, because this choice simplifies computation and provides more stable finite sample performance.

3. Efficient computation. In this section, we develop a computationally efficient Newton–Raphson-type algorithm to minimize the LsoCV score.

3.1. *Shortcut formula.* The definition of LsoCV would indicate that it is necessary to solve n separate minimization problems in order to find the LsoCV score. However, a computational shortcut is available that requires solving only one minimization problem that involves all data. Recall that $\mathbf{A}$ is the hat matrix. Let $\mathbf{A}_{ii}$ denote the diagonal block of $\mathbf{A}$ corresponding to the observations of subject i.

LEMMA 3.1. *(Shortcut Formula) The LsoCV score satisfies*

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-T}(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) \tag{16}$$

where $\mathbf{I}_{ii}$ *is a* $n_i \times n_i$ *identity matrix, and* $\hat{\boldsymbol{y}}_i = \hat{\mu}(\mathbf{X}_i)$.

This result, whose proof is given in the Supplementary Material, extends a similar result for independent data (e.g., Green and Silverman, 1994, page 31). Indeed, if each subject has only one observation, then (16) reduces to LsoCV $= (1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(1-a_{ii})^2$, which is exactly the shortcut formula for the ordinary cross-validation score.

3.2. *An approximation of Leave-subject-out CV.* A close inspection of the short-cut formula of $\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})$ given in (16) suggests that, the evaluation of $\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})$ can still be computationally expensive because of the requirement of matrix inversion and the formulation of the hat matrix $\mathbf{A}$. To further reduce the computational cost, using Taylor's expansion $(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1} \approx \mathbf{I}_{ii} + \mathbf{A}_{ii}$, we obtain the following approximation of $\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})$:

$$\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\mathbf{Y} + \frac{2}{n}\sum_{i=1}^{n}\hat{\boldsymbol{e}}_i^T\mathbf{A}_{ii}\hat{\boldsymbol{e}}_i, \tag{17}$$

where $\hat{\boldsymbol{e}} = (\mathbf{I}-\mathbf{A})\mathbf{Y}$. The next theorem shows that this approximation is a good one in the sense that its minimization is asymptotically equivalent to the minimization of the true loss function.

THEOREM 3.1. *Under Conditions 1–5, for a pre-determined* $\mathbf{W}$ *and a non-random* $\boldsymbol{\lambda}$*, as* $n \to \infty$*, we have*

$$\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - L(\mathbf{W},\boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W},\boldsymbol{\lambda})).$$

This result and Theorem 2.2 together imply that $\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})$ and $\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})$ are asymptotically equivalent, that is, for a pre-determined $\mathbf{W}$ and a non-random $\boldsymbol{\lambda}$, $\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) = o_p(L(\mathbf{W},\boldsymbol{\lambda}))$. The proof of Theorem 3.1 is given in the Appendix.

We developed an efficient algorithm to minimizing $\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ for a pre-given $\mathbf{W}$ based on the works of Gu and Wahba (1991) and Wood (2004). The idea is to optimize the log transform of $\boldsymbol{\lambda}$ using the Newton–Raphson method. The detailed algorithm is described in the Supplementary Material and it can be show that, for $\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})$, the overall computational cost for each Newton–Raphson iteration is $O(Np)$, which is much smaller than the cost of minimizing $\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda})$ ($O(Np^2)$) when the total number of used basis functions p is large.

4. Simulation studies.

4.1. *Function estimation.* In this section, we illustrate the finite-sample performance of LsoCV* in selecting the penalty parameters. In each simulation run, we set $n = 100$ and $n_i = 5$, $i = 1, \cdots, n$. A random sample is generated from the model

$$y_{ij} = f_1(x_{1,i}) + f_2(x_{2,ij}) + \epsilon_{ij}, \qquad j = 1, \cdots, 5,\ i = 1, \cdots, 100, \tag{18}$$

where x_1 is a subject level covariate and x_2 is an observational level covariate, both of which are drawn from $Uniform(-2, 2)$. Functions used here are from Welsh et al. (2002):

$$f_1(x) = \sqrt{z(1-z)} \sin(2\pi \frac{1 + 2^{-3/5}}{1 + z^{-3/5}}),$$

$$f_2(x) = \sin(8z - 4) + 2 \exp(-256(z - 0.5)^2),$$

where $z = (x+2)/4$. The error term ϵ_{ij}'s are generated from a Gaussian distribution with zero mean, variance σ^2, and the compound symmetry within-subject correlation, that is

$$Corr(\epsilon_{ij}, \epsilon_{kl}) = \begin{cases} 1, & \text{if } i = j = k = l; \\ \rho, & \text{if } i = k,\ j \neq l, \\ 0, & otherwise; \end{cases} \tag{19}$$

$j, l = 1, \cdots, 5,\ i, k = 1, \cdots, 100$. In this subsection, we take $\sigma = 1$ and $\rho = 0.8$. A cubic splines with 10 equally spaced interior knots in $[-2, 2]$ was used for estimating each function. Functions were estimated by minimizing (2) with two working correlations: the working independence (denoted as $\mathbf{W}_1 = \mathbf{I}$) and the compound symmetry with $\rho = 0.8$ (denoted as $\mathbf{W}_2$). Penalty parameters were selected by minimizing LsoCV*. The top two panels of Figure 1 show that the biases using $\mathbf{W}_1$ and $\mathbf{W}_2$ are almost the same, which is consistent with the conclusion in Zhu et al. (2008) that the bias of function estimation using regression splines does not depend on the choice of the working correlation. The bottom two panels indicate that using the true correlation structure $\mathbf{W}_2$ yields more efficient function estimation; the message is more clear in the estimation of $f_2(x)$.

4.2. *Comparison with an existing method.* Assuming that the structure of $\mathbf{W}$ is known up to a parameter γ and the true covariance matrix $\mathbf{\Sigma}$ is attained at $\gamma = \gamma_0$, Han and Gu (2008) proposed to simultaneously select γ and $\boldsymbol{\lambda}$ by minimizing the following criterion

$$\mathrm{V}^*(\mathbf{W}, \boldsymbol{\lambda}) = \log\{\mathbf{Y}^T \mathbf{W}^{1/2} (\mathbf{I} - \tilde{\mathbf{A}})^2 \mathbf{W}^{1/2} \mathbf{Y}/N\} - \frac{1}{N} \log |\mathbf{W}| + \frac{2tr(\mathbf{A})}{N - tr(\mathbf{A})}, \tag{20}$$

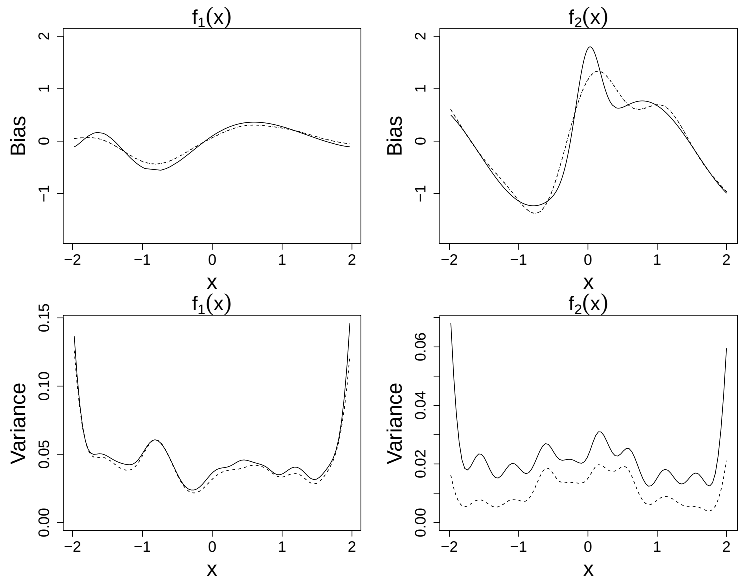

FIG 1. *Simulation results for function estimation based on 200 Monte Carlo runs. Functions are evaluted over* 100 *equally spaced grid points in* $[-2, 2]$. *Top panels: estimated functions: solid—true functions; dashed—average of estimates using* $\mathbf{W}_1$*; dotted—average of estimates using* $\mathbf{W}_2$ *(not distinguishable with dashed). Bottom panels: variance of estimated functions: solid—estimates using* $\mathbf{W}_1$*; dashed—estimates using* $\mathbf{W}_2$.

where N is the total number of observations. They proved that V* is asymptotically optimal in selecting both the penalty parameter $\boldsymbol{\lambda}$ and the correlation parameter γ, provided that the within subject correlation structure is correctly specified. In this section, we compare the finite sample performance of LsoCV* and V* in selecting the penalty parameter when the working correlation matrix $\mathbf{W}$ is given and fixed.

We generated data using (18) and (19) as in the previous subsection and considered different parameters for the correlation matrix. In particular, we fixed $\rho = 0.8$ and varied the noise standard deviation σ from 0.5 to 1, we also fixed $\sigma = 1$ and varied ρ from -0.2 to 0.9. A cubic spline with 10 equally spaced interior knots was used for each unknown regresssion function. For each simulation run, to compare the effectiveness of two selection criteria for a given working correlation matrix $\mathbf{W}$, we calculated the ratio of true losses at different choices of penalty parameters: $L(\mathbf{W}, \boldsymbol{\lambda}_{\mathrm{V}*})/L(\mathbf{W}, \boldsymbol{\lambda}_{\mathrm{LsoCV}*})$ and $L(\mathbf{W}, \boldsymbol{\lambda}_{\mathrm{Opt}})/L(\mathbf{W}, \boldsymbol{\lambda}_{\mathrm{LsoCV}*})$, where $\boldsymbol{\lambda}_{V^*}$ and $\boldsymbol{\lambda}_{\mathrm{LsoCV}^*}$ are penalty parameters selected by using V* and LsoCV*, respectively, and $\boldsymbol{\lambda}_{\mathrm{Opt}}$ is obtained by minimizing the true loss function defined in (7) assuming the mean function $\mu(\cdot)$ is known.

In the first experiment, the true correlation matrix was used as the working correlation matrix, denoted as $\mathbf{W}_1$. This is the case that V* is expected to work well according to Han and Gu (2008). Results in Figures 2 indicate that performances of LsoCV* and V* are comparable for this case regardless of values of σ or ρ. In the second experiment, the working correlation structure was chosen to be different from the true correlation structure. Specifically, the working correlation matrix, denoted as $\mathbf{W}_2$, is a truncated version of (19) where the correlation coefficient between ϵ_{i,j_1} and ϵ_{i,j_2} is set to ρ if $|j_1 - j_2| = 1$ and 0 if $|j_1 - j_2| \geq 2$. Results in Figures 3 show that LsoCV* becomes more effective than V* in terms of minimizing the true loss of estimating the true mean function $\hat{\mu}(\cdot)$ as σ or ρ increases. These results are understandable since V* is applied to a situation that it is not designed for and its asymptotic optimality does not hold. Moreover, from the right two panels of Figures 2 and 3, we see that the minimum value of LsoCV* is reasonably close to the true loss function assuming the knowledge of the true function, as indicated by the conclusion of Theorem 3.1.

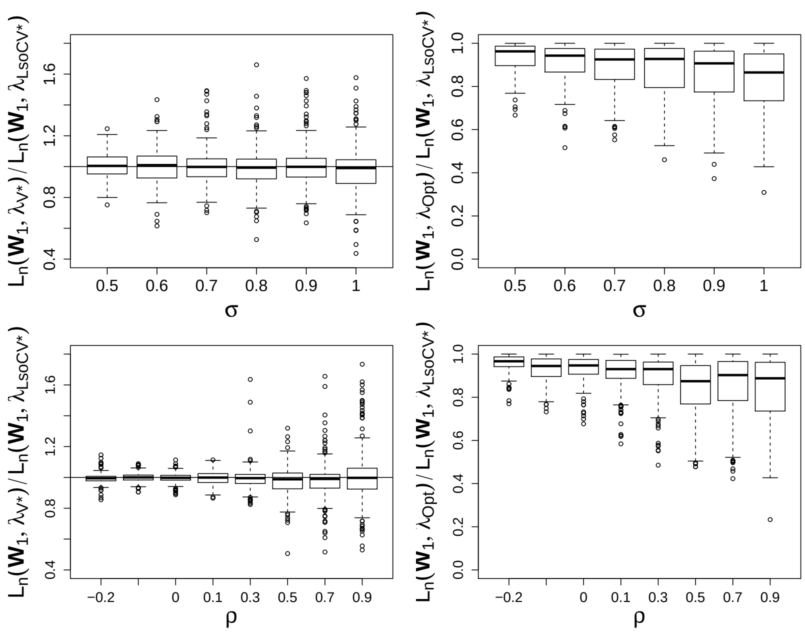

FIG 2. *Relative efficiency of* LsoCV* *to* V* *and to the true loss when the working correlation matrix is the same as the true correlation matrix.*

4.3. *Correlation structure selection.* In this subsection, we study the performance of LsoCV* in selecting the working correlation matrix **W**.

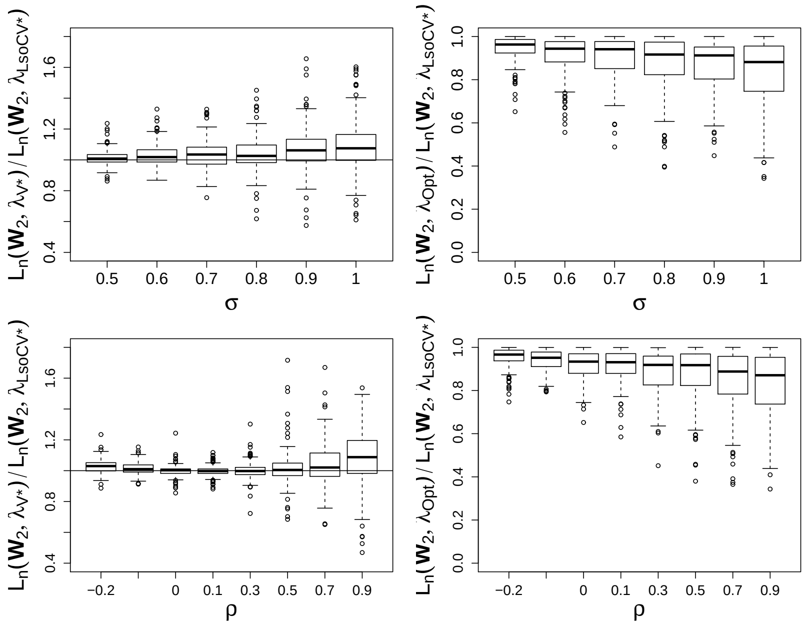

FIG 3. *Relative efficiency of* LsoCV* *to* V* *and to the true loss when the working correlation matrix is different from the true correlation matrix.*

The data was generated using the model (18) with $\sigma = 1$, $n_i = 5$ for all $i = 1, \cdots, n$. In this experiment, both x_1 and x_2 are set to be observational level covariates drawn from $Uniform(-2, 2)$. Four types of within-subject correlation structures were considered: independence (IND), compound symmetry with correlation coefficient ρ (CS), AR(1) with lag-one correlation ρ (AR), and unstructured correlation matrix with $\rho_{12} = \rho_{23} = 0.8$, $\rho_{13} = 0.3$ and 0 otherwise (UN). Data were generated using one of these correlation structures and then the LsoCV* was used to select the best working correlation from the four possible candidates. A cubic spline with 10 equally spaced interior knots in $[-2, 2]$ was used to model each unknown function and we set the penalty parameter vector $\boldsymbol{\lambda} = \mathbf{0}$. Simulation results based on 200 runs were summarized in Table 1, which show fairly good selection results in the sense that the true correlation structure was selected in majority of times.

5. A real data example. As a subset from the Multi-center AIDS Cohort Study, the data set includes the repeated measurements of CD4 cell counts and percentages on 283 homosexual men who became HIV-positive between 1984 and 1991. All subjects were scheduled to take their measure-

TABLE 1
Simulation results for working correlation structure selection.

			Selected Structure			
n	ρ	True Structure	IND	CS	AR	UN
50	0.3	IND	97.0	2.0	1.0	0
		CS	8.5	78.0	13.5	0
		AR	13.5	10.0	76.5	0
		UN	1.5	1.5	21.5	75.5
	0.5	IND	96.5	2.5	1.0	0
		CS	3.0	78.5	18.5	0
		AR	4.0	9.5	86.5	0
		UN	3.5	4.0	11.5	81.0
	0.8	IND	98.5	1.0	0.5	0
		CS	3.5	74.0	22.0	0.5
		AR	5.5	21.0	71.0	2.5
		UN	5.5	1.0	8.5	85.0
100	0.3	IND	95.0	3.0	2.0	0
		CS	2.0	84.5	13.5	0
		AR	3.5	8.5	88.0	0
		UN	0	1.0	13.5	85.5
	0.5	IND	99.5	0.5	0	0
		CS	2.5	81.0	16.5	0
		AR	1.0	6.0	93.0	0
		UN	2.0	0.5	10.0	87.5
	0.8	IND	99.0	1.0	0	0
		CS	2.5	73.5	24.0	0
		AR	2.0	20.0	76.5	1.5
		UN	5.5	2.0	9.0	83.5
150	0.3	IND	98.5	1.0	0.5	0
		CS	2.0	85.0	13.0	0
		AR	2.5	5.5	92.0	0
		UN	0	0	16.5	83.5
	0.5	IND	100	0	0	0
		CS	1.0	81.5	17.5	0
		AR	2.5	8.5	89.0	0
		UN	0.5	0	12.0	87.5
	0.8	IND	99.5	0.5	0	0
		CS	1.0	78.0	20.0	1.0
		AR	0.5	18.5	77.5	3.5
		UN	1.0	2.0	6.5	90.5

ments at semi-annual visits. However, since many subjects missed some of their scheduled visits, there are unequal numbers of repeated measurements and different measurement times per subject. Further details of the study can be found in Kaslow et al. (1987).

Our goal is to do statistical analysis of the trend of mean CD4 percentage depletion over time. Denote by t_{ij} the time in years of the jth measurement of the ith individual after HIV infection, by y_{ij} the ith individual's CD4 percentage at time t_{ij} and by $X_i^{(1)}$ the ith individual's smoking status with values 1 or 0 for the ith individual ever or never smoked cigarettes, respectively, after the HIV infection. To obtain a clear biological interpretation, we define $X_i^{(2)}$ to be the ith individual's centered age at HIV infection, which is obtained by the ith individual's age at infection subtract the sample average age at infection. Similarly, the ith individual's centered pre-infection CD4 percentage, denoted by $X_i^{(3)}$, is computed by subtracting the average pre-infection CD4 percentage of the sample from the ith individual's actual pre-infection CD4 percentage. These covariates, except the time, are time-invariant. Consider the varying-coefficient model

$$y_{ij} = \beta_0(t_{ij}) + X_i^{(1)}\beta_1(t_{ij}) + X_i^{(2)}\beta_2(t_{ij}) + X_i^{(2)}\beta_2(t_{ij}), \tag{21}$$

where $\beta_0(t)$ represents the trend of mean CD4 percentage changing over time after the infection for a non-smoker with average pre-infection CD4 percentage and average age at HIV infection, and $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ describe the time-varying effects for cigarette smoking, age at HIV infection, and pre-infection CD4 percentage, respectively, on the post-infection CD4 percentage. Since the number observations are very uneven among subjects, we only used subjects with at least 4 observations. A cubic spline with $k = 10$ equally spaced knots was used for modeling each function. We first used the working independence $\mathbf{W}_1 = \mathbf{I}$ to fit the data and then use the residuals from this model to estimate parameters in the correlation function

$$\gamma(u; \alpha, \theta) = \alpha + (1 - \alpha)\exp(-\theta u),$$

where u is the lag in time and $0 < \alpha < 1$, $\theta > 0$. This correlation function was considered previously in Zeger and Diggle (1994). The estimated parameter values are $(\hat{\alpha}, \hat{\theta}) = (0.40, 0.75)$. The second working correlation matrix $\mathbf{W}_2$ considered was formed using $\gamma(u; \hat{\alpha}, \hat{\theta})$. We computed that $\text{LsoCV}(\mathbf{W}_1, \mathbf{0}) = 881.88$ and $\text{LsoCV}(\mathbf{W}_2, \mathbf{0}) = 880.33$, which implies that using $\mathbf{W}_2$ may be more desirable. This conclusion remains unchanged when the number of knots varies. To visualize the gain in estimation efficiency by using $\mathbf{W}_2$ instead of $\mathbf{W}_1$, we calculated the width of the 95% pointwise bootstrap confidence intervals based on 1000 bootstrap samples, which is displayed in Figure 4. We can observe that the bootstrap intervals using $\mathbf{W}_2$ is almost uniformly narrower than those using $\mathbf{W}_1$, indicating higher estimation efficiency. The fitted coefficient functions (not shown to save space) using $\mathbf{W}_2$ with $\boldsymbol{\lambda}$ selected by minimizing LsoCV*$(\mathbf{W}_2, \boldsymbol{\lambda})$ are similar to those published in previous studies conducted on the same data set (Wu and Chiang, 2000; Fan and Zhang, 2000; Huang et al., 2002).

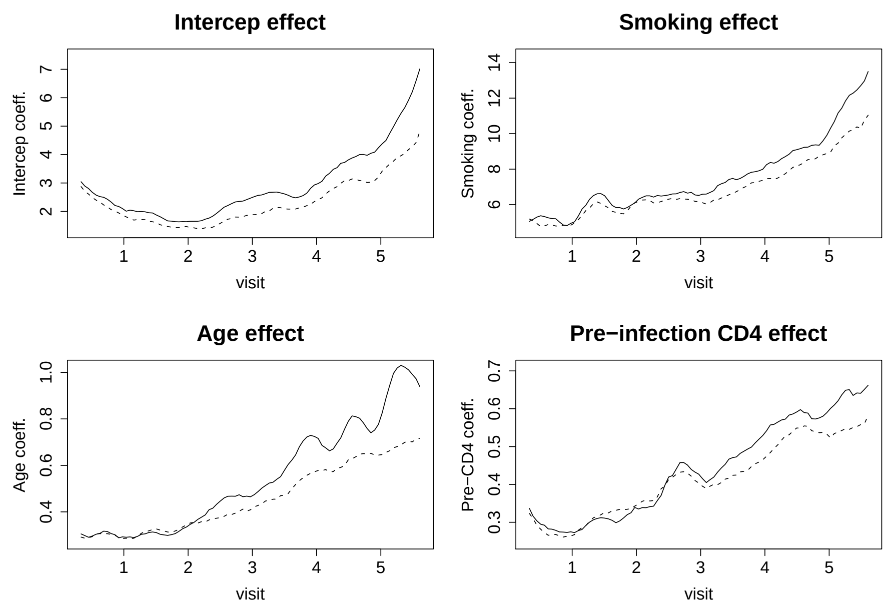

FIG 4. *Width of the* 95% *pointwise bootstrap confidence intervals based on 1000 bootstrap samples, using the working independence* $\mathbf{W}_1$ *(solid line) and the working correlation matrix* $\mathbf{W}_2$ *(dashed line).*

APPENDIX A: TECHNICAL PROOFS

This section is organized as follows. We first give three technical lemmas (Lemma A.1-A.4) needed for the proof of Theorem 2.1. After proving Theorem 2.1, we give another lemma (Lemma A.5) that facilitates proofs of Theorem 2.2 and 3.1. We prove Theorem 3.1 first and then proceed to the proof of Theorem 2.2.

Let $\lambda_{max}(\mathbf{M}) = \lambda_1(\mathbf{M}) \geq \lambda_2(\mathbf{M}) \geq \cdots \geq \lambda_p(\mathbf{M}) = \lambda_{min}(\mathbf{M})$ be eigenvalues of the $p \times p$ symmetric matrix $\mathbf{M}$. We present several useful lemmas.

LEMMA A.1. *For any positive semi-definite matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$*,*

$$\lambda_i(\mathbf{M}_1)\lambda_p(\mathbf{M}_2) \leq \lambda_i(\mathbf{M}_1\mathbf{M}_2) \leq \lambda_i(\mathbf{M}_1)\lambda_1(\mathbf{M}_2), \quad i = 1, \cdots, p\,. \tag{22}$$

Proof. See Anderson and Gupta (1963) and Benasseni (2002). □

LEMMA A.2. *For any positive semi-definite matrix* $\mathbf{M}_1$*, and* $\mathbf{M}_2$*,*

$$tr(\mathbf{M}_1\mathbf{M}_2) \leq \lambda_{max}(\mathbf{M}_1)tr(\mathbf{M}_2), \tag{23}$$

Proof. The proof is trivial, using the eigen decomposition of $\mathbf{M}_1$. □

LEMMA A.3. *Eigenvalues of* $\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}$ *and* $(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\mathbf{\Sigma}$ *are bounded above by* $\xi(\mathbf{\Sigma}, \mathbf{W}) = \lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})$*.*

Proof. Recall that $\tilde{\mathbf{A}} = \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{1/2}$. For $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$, by Lemma A.1,

$$\begin{aligned}\lambda_i(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}) &= \lambda_i(\tilde{\mathbf{A}}\mathbf{W}\tilde{\mathbf{A}}\mathbf{W}^{-1/2}\mathbf{\Sigma}\mathbf{W}^{-1/2}) \\ &\leq \lambda_i(\tilde{\mathbf{A}}\mathbf{W}\tilde{\mathbf{A}})\lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1}) \\ &\leq \lambda_i(\tilde{\mathbf{A}}^2)\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1}) \leq \xi(\mathbf{\Sigma}, \mathbf{W}).\end{aligned}$$

The last inequality follows from the fact that $\max_i\{\lambda_i(\tilde{\mathbf{A}}^2)\} \leq 1$. Similarly, $\lambda_i((\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\mathbf{\Sigma}) \leq \xi(\mathbf{\Sigma}, \mathbf{W})$ follows from $\max_i\{\lambda_i((\mathbf{I}-\tilde{\mathbf{A}})^2)\} \leq 1$. □

Denote $\boldsymbol{e} = (\boldsymbol{e}_1^T, \cdots, \boldsymbol{e}_n^T)^T$, where $\boldsymbol{e}_i$'s are independent random vectors with length n_i, $E(\boldsymbol{e}_i) = 0$ and $Var(\boldsymbol{e}) = \mathbf{I}_i$ for $i = 1, \cdots, n$. For each i, define $z_{ij} = (\boldsymbol{u}_{ij}^T\boldsymbol{e}_i)^2$ where $\boldsymbol{u}_{ij}^T\boldsymbol{u}_{ik} = 1$ if $j = k$ and 0 otherwise, $j, k = 1, \cdots, n_i$.

LEMMA A.4. *If there exists a constant* K *such that* $E(z_{ij}^2) \leq K$ *holds for all* $j = 1, \cdots, n_i$*,* $i = 1, \cdots, n$*, then*

$$Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) \leq 2tr(\mathbf{B}\mathbf{B}^T) + K\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^*)\}^2, \tag{24}$$

where $\mathbf{B}$ *is any* $N \times N$ *matrix (not necessarily symmetric),* $\mathbf{B}_{ii}$ *is the ith* $(n_i \times n_i)$ *diagonal block of* $\mathbf{B}$ *and* $\mathbf{B}_{ii}^*$ *is an "envelop" matrix such that* $\mathbf{B}_{ii}^* \pm (\mathbf{B}_{ii} + \mathbf{B}_{ii}^T)/2$ *are positive semi-definite.*

The proof is given in the Supplementary Material.

Proof of Theorem 2.1. In light of (9) and (15), it suffices to show that

$$L(\mathbf{W},\boldsymbol{\lambda}) - R(\mathbf{W},\boldsymbol{\lambda}) = o_p(R(\mathbf{W},\boldsymbol{\lambda})), \tag{25}$$

$$\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\boldsymbol{\epsilon} = o_p(R(\mathbf{W},\boldsymbol{\lambda})), \tag{26}$$

$$\frac{2}{n}\{\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon} - tr(\mathbf{A}\boldsymbol{\Sigma})\} = o_p(R(\mathbf{W},\boldsymbol{\lambda})) \tag{27}$$

because, combining (25)–(27), we have

$$U(\mathbf{W},\boldsymbol{\lambda}) - L(\mathbf{W},\boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W},\boldsymbol{\lambda})).$$

We first prove (25). By (8), we have

$$Var(L(\mathbf{W},\boldsymbol{\lambda})) = \frac{1}{n^2}Var\{\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon} - 2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}\boldsymbol{\epsilon}\}. \tag{28}$$

Define $\mathbf{B} = \boldsymbol{\Sigma}^{1/2}\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}^{1/2}$. Then $\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon} = (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon})^T\mathbf{B}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon})$. Since $\mathbf{B}$ is positive semi-definite, by applying Lemma A.4 with $\boldsymbol{e} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$, $\mathbf{B} = \boldsymbol{\Sigma}^{1/2}\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}^{1/2}$ and $\mathbf{B}^*_{ii} = \mathbf{B}_{ii}$, we obtain

$$\frac{1}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon}) \le \frac{2}{n^2}tr(\mathbf{B}^2) + \frac{K}{n^2}\sum_{i=1}^n\{tr(\mathbf{B}_{ii})\}^2, \tag{29}$$

for some $K > 0$ as defined in Lemma A.4. By Lemma A.2 and Lemma A.3, under Condition 3, we have

$$\begin{aligned}\frac{2}{n^2}tr(\mathbf{B}^2) &\le \frac{2\lambda_{max}(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma})}{n^2}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma})\\ &\le \frac{2\xi(\boldsymbol{\Sigma},\mathbf{W})}{n}\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}) = o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned} \tag{30}$$

Reall $\mathbf{C}_{ii}$ is the ith diagonal block of $\tilde{\mathbf{A}}^2$. Then, under Condition 2(ii), $tr(\mathbf{C}_{ii}) \sim o(1)$. Thus,

$$\begin{aligned}tr(\mathbf{B}_{ii}) &= tr(\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1/2}\tilde{\mathbf{A}}\mathbf{W}\tilde{\mathbf{A}}\mathbf{W}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T)\\ &\le \lambda_{max}(\mathbf{W})tr(\tilde{\mathbf{A}}\mathbf{W}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1/2}\tilde{\mathbf{A}})\\ &= \lambda_{max}(\mathbf{W})tr(\mathbf{C}_{ii}\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1/2})\\ &\le \lambda_{max}(\mathbf{W})\lambda_{max}(\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})tr(\mathbf{C}_{ii})\\ &= o(1)\xi(\boldsymbol{\Sigma},\mathbf{W}).\end{aligned} \tag{31}$$

Since $\sum_{i=1}^n \{tr(\mathbf{B}_{ii})\} = tr(\mathbf{B}) = tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma})$, under Condition 3,

$$\begin{aligned}(32)\qquad \frac{K}{n^2}\sum_{i=1}^n\{tr(\mathbf{B}_{ii})\}^2 &= o(1)\frac{K\xi(\mathbf{\Sigma},\mathbf{W})tr(\mathbf{B})}{n^2}\\ &= o(1)\frac{K\xi(\mathbf{\Sigma},\mathbf{W})}{n}\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}) = o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

Combining (29)–(32), we obtain

$$\frac{1}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon}) \sim o(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

Since $\lambda_{max}(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}) \le \xi(\mathbf{\Sigma},\mathbf{W})$ by Lemma A.3, under Condition 3,

$$\begin{aligned}(33)\qquad \frac{1}{n^2}Var\{\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}\boldsymbol{\epsilon}\} &= \frac{1}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &\le \frac{\lambda_{max}(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &\le \frac{\xi(\mathbf{\Sigma},\mathbf{W})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &= o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

Combining (28)-(33) and using the Cauchy–Schwarz inequality, we obtain $Var(L(\mathbf{W},\boldsymbol{\lambda})) = o(R^2(\mathbf{W},\boldsymbol{\lambda}))$, which proves (25).

To show (26), by Lemma (A.3) and Condition 3, we have

$$\begin{aligned}\frac{1}{n^2}Var\{\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\boldsymbol{\epsilon}\} &= \frac{1}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{\Sigma}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &\le \frac{\lambda_{max}(\mathbf{\Sigma})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &\le \frac{\xi(\mathbf{\Sigma},\mathbf{W})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

The result follows from an application of the Chebyshev inequality.

To show (27), applying Lemma A.4 with $\boldsymbol{e} = \mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$, $\mathbf{B} = \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$. For each $\mathbf{B}_{ii} = \mathbf{\Sigma}_i^{1/2}\mathbf{A}_{ii}\mathbf{\Sigma}_i^{1/2}$, noticing that $(\mathbf{W}_i^{1/2} - \alpha\mathbf{W}_i^{-1/2})\tilde{\mathbf{A}}_{ii}(\mathbf{W}_i^{1/2} - \alpha\mathbf{W}_i^{-1/2})$ is positive semi-definite, we can define an "envelop" matrix as $\mathbf{B}_{ii}^* = \frac{1}{2}\mathbf{\Sigma}_i^{1/2}(\mathbf{W}_i^{1/2}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{1/2}/\alpha_i + \alpha_i\mathbf{W}_i^{-1/2}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{-1/2})\mathbf{\Sigma}_i^{1/2}$ for any $\alpha_i > 0$. Then by Lemma A.4, we obtain

$$\begin{aligned}(34)\qquad \frac{2}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}) &= \frac{2}{n^2}Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e})\\ &\le \frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) + \frac{K}{n^2}\sum_{i=1}^n\{tr(\mathbf{B}_{ii}^*)\}^2,\end{aligned}$$

where K is as in Lemma A.4. By Lemma A.2, under Condition 3, we have

$$\frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) = \frac{2}{n^2}tr(\mathbf{\Sigma}\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T) \le \frac{2\lambda_{max}(\mathbf{\Sigma})}{n}\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma})$$
$$\le \frac{2\xi(\mathbf{\Sigma},\mathbf{W})}{n}\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}) = o(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

By using Lemma A.1 repeatedly and taking $\alpha_i = \lambda_{max}(\mathbf{W}_i)$, we have

$$tr(\mathbf{B}^*_{ii}) = tr(\tilde{\mathbf{A}}_{ii}\mathbf{\Sigma}_i^{1/2}\mathbf{W}_i\mathbf{\Sigma}_i^{1/2})/(2\alpha_i) + \alpha_i tr(\tilde{\mathbf{A}}_{ii}\mathbf{\Sigma}_i^{1/2}\mathbf{W}_i^{-1}\mathbf{\Sigma}_i^{1/2})/2$$
$$\le \lambda_{max}(\mathbf{\Sigma}_i\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{A}}_{ii})$$
$$\le \xi(\mathbf{\Sigma},\mathbf{W})tr(\tilde{\mathbf{A}}_{ii}).$$

Under Conditions 2(i), 3 and 4, we have

$$\frac{K}{n^2}\sum_{i=1}^{n}\{tr(\mathbf{B}^*_{ii})\}^2 \le \frac{K}{n^2}\xi^2(\mathbf{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{35}$$

Therefore, combining (34)–(35) and noticing Conditions 1–4, we have

$$\frac{1}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}) \sim o(R^2(\mathbf{W},\boldsymbol{\lambda})),$$

which leads to (27). □

To prove Theorem 2.2, it is easier to prove Theorem 3.1 first. The following lemma is useful for the proof of Theorem 3.1.

LEMMA A.5. *Let $\mathbf{D} = diag\{\mathbf{D}_{11},\cdots,\mathbf{D}_{nn}\}$ be a diagonal block matrix and $\mathbf{D}^* = diag\{\mathbf{D}^*_{11},\cdots,\mathbf{D}^*_{nn}\}$ be a positive semi-definite matrix such that $\mathbf{D}^* \pm (\mathbf{D}+\mathbf{D}^T)/2$ are positive semi-definte. In addition, $\mathbf{D}_{ii}$'s and $\mathbf{D}^*_{ii}$'s meet conditions: (i) $\max_{1\le i\le n}\{tr(\mathbf{D}^*_{ii}\mathbf{W}_i)\} \sim \lambda_{max}(\mathbf{W})O(n^{-1}tr(\mathbf{A}))$; (ii) $\max_{1\le i\le n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii}^T\} \sim \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$. Then, under Conditions 1–5, we have*

$$\frac{1}{n^2}Var\{\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\mathbf{Y}\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

The proof is given in the Supplementary Material.

Proof of Theorem 3.1. By Theorem 2.1, it suffices to show that

$$\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda}) = o_p(R(\mathbf{W},\boldsymbol{\lambda})),$$

which can be obtained by showing

$$E\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\}^2 = o_p(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{36}$$

Hence, it suffices to show that

$$E\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\} = o(R(\mathbf{W},\boldsymbol{\lambda})), \quad \text{and} \tag{37}$$

$$Var\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{38}$$

Denote $\mathbf{A}_d = diag\{\mathbf{A}_{11},\cdots,\mathbf{A}_{nn}\}$ and $\tilde{\mathbf{A}}_d = diag\{\tilde{\mathbf{A}}_{11},\cdots,\tilde{\mathbf{A}}_{nn}\}$. It follows that $\tilde{\mathbf{A}}_d = \mathbf{W}^{-1/2}\mathbf{A}_d\mathbf{W}^{1/2}$ and $n^{-1}tr(\tilde{\mathbf{A}}_d^2) = O(n^{-2}tr(\mathbf{A})^2)$ by Condition 2. Some algebra yields that

$$\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda}) = \frac{2}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}_d(\mathbf{I}-\mathbf{A})\mathbf{Y} - \frac{2}{n}tr(\mathbf{A}\boldsymbol{\Sigma}).$$

First consider (37). We have that

$$\begin{aligned} &E\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\} \\ &= \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{A}_d+\mathbf{A}_d^T)(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\ &\quad + \frac{1}{n}tr\{\mathbf{A}^T(\mathbf{A}_d+\mathbf{A}_d^T)\mathbf{A}\boldsymbol{\Sigma}\} - \frac{2}{n}tr(\mathbf{A}_d^T\mathbf{A}_d\boldsymbol{\Sigma}) - \frac{2}{n}tr(\mathbf{A}_d^2\boldsymbol{\Sigma}). \end{aligned} \tag{39}$$

We shall show that each term in (39) is of the order $o(R(\mathbf{W},\boldsymbol{\lambda}))$.

Condition 2 says that $\max_{1\le i\le n} tr(\tilde{\mathbf{A}}_{ii}) = O(n^{-1}tr(\mathbf{A})) = o(1)$. Using Conditions 2 and 5, we have

$$\begin{aligned} tr(\mathbf{A}_{ii}+\mathbf{A}_{ii}^T)^2 &= 2tr(\mathbf{A}_{ii}^2 + \mathbf{A}_{ii}\mathbf{A}_{ii}^T) \\ &= 2tr(\tilde{\mathbf{A}}_{ii}^2 + \tilde{\mathbf{A}}_{ii}\mathbf{W}_i\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{-1}) \\ &\le 2tr(\tilde{\mathbf{A}}_{ii}^2)\{1+\lambda_{max}(\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)\} \\ &= \lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2) = o(1), \end{aligned}$$

which implies that all eigenvalues of $(\mathbf{A}_d+\mathbf{A}_d^T)$ are of order $o(1)$, and hence

$$\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{A}_d+\mathbf{A}_d^T)(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(1)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R(\mathbf{W},\boldsymbol{\lambda})),$$

$$\frac{1}{n}tr\{\mathbf{A}^T(\mathbf{A}_d+\mathbf{A}_d^T)\mathbf{A}\boldsymbol{\Sigma}\} = o(1)\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}) = o(R(\mathbf{W},\boldsymbol{\lambda})).$$

Under Condition 4, the third term in (39) can be bounded as

$$\begin{aligned}\frac{1}{n}tr(\mathbf{A}_d^T\mathbf{A}_d\mathbf{\Sigma}) &\le \lambda_{max}(\mathbf{\Sigma W}^{-1})\frac{1}{n}tr(\tilde{\mathbf{A}}_d^{1/2}\mathbf{W}^{1/2}\tilde{\mathbf{A}}_d)\\ &\le \xi(\mathbf{\Sigma},\mathbf{W})\frac{1}{n}tr(\tilde{\mathbf{A}}_d^2)\\ &= \xi(\mathbf{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R(\mathbf{W},\boldsymbol{\lambda})).\end{aligned} \tag{40}$$

For the last term in equation (39), observe that $(\mathbf{W}_i^{1/2}-\alpha_i\mathbf{W}_i^{-1/2})\mathbf{\Sigma}(\mathbf{W}_i^{1/2}-\alpha_i\mathbf{W}_i^{-1/2})$ is positive semi-definite for any α_i. Taking $\alpha_i=\lambda_{max}(\mathbf{W}_i)$, we have

$$\begin{aligned}\frac{2}{n}tr(\mathbf{A}_d^2\mathbf{\Sigma}) = \frac{2}{n}tr(\tilde{\mathbf{A}}_d^2\mathbf{W}^{-1/2}\mathbf{\Sigma W}^{1/2}) &\le \max_{1\le i\le n} tr\{\tilde{\mathbf{A}}_{ii}^2(\mathbf{\Sigma}_i^* + \mathbf{\Sigma}_i^{*T})\}\\ &\le \max_{1\le i\le n} tr\{\tilde{\mathbf{A}}_{ii}^2(\mathbf{W}_i^{1/2}\mathbf{\Sigma}_i\mathbf{W}_i^{1/2}/\alpha_i + \alpha_i\mathbf{W}_i^{-1/2}\mathbf{\Sigma}_i\mathbf{W}_i^{-1/2})\}\\ &\le \max_{1\le i\le n}\{\lambda_{max}(\mathbf{\Sigma}_i\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{A}}_{ii}^2)\}\\ &\le \xi(\mathbf{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R(\mathbf{W},\boldsymbol{\lambda})),\end{aligned}$$

where $\mathbf{\Sigma}_i^* = \mathbf{W}_i^{-1/2}\mathbf{\Sigma}_i\mathbf{W}_i^{1/2}$. Equation (39) and thus (37) have been proved.

To prove (38), define $\mathbf{D} = \mathbf{A}_d$ and the corresponding "envelop" matrix $\mathbf{D}^* = diag\{\mathbf{D}_{11}^*,\cdots,\mathbf{D}_{nn}^*\}$, where the diagonal blocks are defined as $\mathbf{D}_{ii}^* = \frac{1}{2}(\mathbf{W}^{1/2}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{1/2}/\alpha_i + \alpha_i\mathbf{W}_i^{-1/2}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{-1/2})$ with $\alpha_i = \lambda_{max}(\mathbf{W}_i)$, then since

$$tr(\mathbf{A}_{ii}\mathbf{W}_i\mathbf{A}_{ii}^T) = tr(\tilde{\mathbf{A}}_{ii}^2\mathbf{W}_i) \le \lambda_{max}(\mathbf{W}_i)\{tr(\mathbf{A}_{ii})\}^2, \quad \text{and}$$

$$tr(\mathbf{D}_{ii}^*\mathbf{W}_i) \le \lambda_{max}(\mathbf{W}_i)tr(\mathbf{A}_{ii}),$$

we have that $\max_{1\le i\le n} tr(\mathbf{A}_{ii}\mathbf{W}_i\mathbf{A}_{ii}^T) = \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$ and that $\max_{1\le i\le n} tr(\mathbf{D}_{ii}^*\mathbf{W}_i) = \lambda_{max}(\mathbf{W})O(n^{-1}tr(\mathbf{A}))$ by Condition 2. Under Conditions 3–4, (38) follows from Lemma A.5. □

Proof of Theorem 2.2. By Theorem 3.1, it suffices to show

$$\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) = o_p(L(\mathbf{W},\boldsymbol{\lambda})),$$

which can be proved by showing that

$$E\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\}^2 = o_p(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

It suffices to show

$$E\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\} = o(R(\mathbf{W},\boldsymbol{\lambda})) \quad \text{and,} \tag{41}$$

$$Var\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{42}$$

For each $i = 1,\ldots,n$, consider the eigen-decomposition $\tilde{\mathbf{A}}_{ii} = \mathbf{P}_i\boldsymbol{\Lambda}_i\mathbf{P}_i^T$, where $\mathbf{P}_i$ is a $n_i \times n_i$ orthogonal matrix and $\boldsymbol{\Lambda}_i = diag\{\lambda_{i1},\cdots,\lambda_{in_i}\}$, $\lambda_{ij} \geq 0$. Using this decomposition, we have

$$(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1} = \mathbf{W}_i^{1/2}\mathbf{P}_i\boldsymbol{\Lambda}_i^*\mathbf{P}_i^T\mathbf{W}^{-1/2},$$

where $\boldsymbol{\Lambda}_i^*$ is a diagonal matrix with diagonal elements $(1-\lambda_{ij})^{-1}$, $j = 1,\cdots,n_i$. Since under Condition 2, $\max_{1\leq j\leq n_i}\{\lambda_{ij}\} \sim o(1)$, we have $(1-\lambda_{ij})^{-1} = \sum_{k=0}^{\infty}\lambda_{ij}^k$, which leads to

$$(\mathbf{I}_{ii} - \tilde{\mathbf{A}}_{ii})^{-1} = \sum_{k=0}^{\infty}\mathbf{P}_i\boldsymbol{\Lambda}_i^k\mathbf{P}_i^T = \sum_{k=0}^{\infty}\tilde{\mathbf{A}}_{ii}^k.$$

Define $\tilde{\mathbf{D}}^{(m)} = diag\{\tilde{\mathbf{D}}_{11}^{(m)},\cdots,\tilde{\mathbf{D}}_{nn}^{(m)}\}$, where $\tilde{\mathbf{D}}_{ii}^{(m)} = \sum_{k=m}^{\infty}\tilde{\mathbf{A}}_{ii}^k$ $i = 1,\cdots,n$, $m = 1,2,\ldots$. It follows that, for each i,

$$tr(\tilde{\mathbf{D}}_{ii}^{(m)}) = \sum_{k=m}^{\infty}tr(\tilde{\mathbf{A}}_{ii}^k) \leq \sum_{k=m}^{\infty}\{tr(\tilde{\mathbf{A}}_{ii})\}^k = \frac{\{tr(\tilde{\mathbf{A}}_{ii})\}^m}{1-tr(\tilde{\mathbf{A}}_{ii})}.$$

Since Condition 2(i) gives $\max_{1\leq i\leq n}tr(\mathbf{A}_{ii}) \sim O(n^{-1}tr(\mathbf{A}))$, we obtain that

$$\max_{1\leq i\leq n}tr(\tilde{\mathbf{D}}_{ii}^{(m)}) = O(n^{-m}tr(\mathbf{A})^m), \quad m = 1,2,\ldots. \tag{43}$$

Some algebra yields

$$\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{D}^{(1)}+\mathbf{D}^{(2)})^{1/2}(\mathbf{I}-\mathbf{A})\mathbf{Y}$$

where $\mathbf{D}^{(1)} = \mathbf{W}^{-1/2}\tilde{\mathbf{D}}^{(1)}\mathbf{W}\tilde{\mathbf{D}}^{(1)}\mathbf{W}^{-1/2}$ and $\mathbf{D}^{(2)} = \mathbf{W}^{1/2}\tilde{\mathbf{D}}^{(2)}\mathbf{W}^{-1/2}$.

To show (41), note that

$$\begin{aligned} &E\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\} \\ &= \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\Sigma\} \\ &\quad + \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\Sigma\}. \end{aligned} \tag{44}$$

Using Lemma A.1 and A.2 repeatedly and Condition 5, we have

$$\lambda_{max}(\mathbf{D}^{(1)}) \leq \lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2) = o(1).$$

Thus, the first terms (44) can be bounded as

$$\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(1)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R(\mathbf{W},\boldsymbol{\lambda})).$$

Using Lemma A.3, under Condition 4 and (43), the second term of (44) can be bounded as

$$\begin{aligned}\frac{1}{n}tr\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\Sigma\} &\le \xi(\boldsymbol{\Sigma},\mathbf{W})\frac{1}{n}tr(\tilde{\mathbf{D}}^{(1)2})\\ &= \xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

Now consider the third term in (44). Under Condition 5 and (43),

$$\begin{aligned} tr\{(\mathbf{D}_{ii}^{(2)}+\mathbf{D}_{ii}^{(2)T})^2\} &= 2tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) + 2tr(\mathbf{D}_{ii}^{(2)}\mathbf{D}_{ii}^{(2)T})\\ &= 2tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) + 2tr(\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{-1}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i)\\ &\le 2tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) + 2\lambda_{max}(\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{D}}_{ii}^{(2)2})\\ &= o(n^{-2}tr(\mathbf{A})^2),\end{aligned} \tag{45}$$

which implies that all eigenvalues of $\mathbf{D}_{ii}^{(2)}+\mathbf{D}_{ii}^{(2)T}$ are of the order $O(n^{-1}tr(\mathbf{A}))$, and thus $o(1)$. Then, under Conditions 1–5, we have

$$\begin{aligned}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} &= \frac{1}{2n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{D}^{(2)}+\mathbf{D}^{(2)T})(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &= o(1)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

To study the the fourth term in (44), we have

$$\begin{aligned}\frac{1}{n}tr\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}\} &= \frac{1}{n}\sum_{i=1}^n tr\{(\mathbf{I}_{ii}-\mathbf{A}_{ii})^T\mathbf{D}_{ii}^{(2)}(\mathbf{I}_{ii}-\mathbf{A}_{ii})\boldsymbol{\Sigma}_i\}\\ &\quad -\frac{1}{n}\sum_{i=1}^n tr(\mathbf{A}_{ii}^T\mathbf{D}_{ii}^{(2)}\mathbf{A}_{ii}\boldsymbol{\Sigma}_i) + \frac{1}{n}tr(\mathbf{A}^T\mathbf{D}^{(2)}\mathbf{A}\boldsymbol{\Sigma}).\end{aligned} \tag{46}$$

To bound the first term in (46), we note that

$$\begin{aligned}& tr\{(\mathbf{I}_{ii}-\mathbf{A}_{ii})^T\mathbf{D}_{ii}^{(2)}(\mathbf{I}_{ii}-\mathbf{A}_{ii})\boldsymbol{\Sigma}_i\}\\ &= \frac{1}{2}tr\{(\mathbf{I}_{ii}-\mathbf{A}_{ii})^T(\mathbf{W}_i^{1/2}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{-1/2}+\mathbf{W}_i^{-1/2}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{1/2})(\mathbf{I}_{ii}-\mathbf{A}_{ii})\boldsymbol{\Sigma}_i\},\end{aligned}$$

which is bounded by

$$\frac{1}{2}tr\{(\mathbf{I}_{ii}-\mathbf{A}_{ii})^T(\mathbf{W}_i^{1/2}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{1/2}/\alpha_i+\alpha_i\mathbf{W}_i^{-1/2}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{-1/2})(\mathbf{I}_{ii}-\mathbf{A}_{ii})\mathbf{\Sigma}_i\}$$
$$\leq \frac{1}{2}\xi(\mathbf{\Sigma}_i,\mathbf{W}_i)tr(\tilde{\mathbf{D}}_{ii}^{(2)})+\frac{\alpha_i}{2}tr\{(\tilde{\mathbf{D}}_{ii}^{(2)}-2\tilde{\mathbf{D}}_{ii}^{(3)})\mathbf{W}_i^{-1/2}\mathbf{\Sigma}_i\mathbf{W}_i^{-1/2}\}$$
$$+\frac{\alpha_i}{2}tr\{\tilde{\mathbf{D}}_{ii}^{(2)}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{-1/2}\mathbf{\Sigma}_i\mathbf{W}_i^{-1/2}\tilde{\mathbf{A}}_{ii}\}$$
$$\leq \frac{1}{2}\xi(\mathbf{\Sigma},\mathbf{W})\{2+\lambda_{max}(\tilde{\mathbf{A}}_{ii}^2)\}tr(\tilde{\mathbf{D}}_{ii}^{(2)})$$
$$=o(R(\mathbf{W},\boldsymbol{\lambda})),$$

where we take $\alpha_i=\lambda_{max}(\mathbf{W}_i)$. The last equation follows from (43) and Condition 4. Similarly, we can show that the second part of (46) is $o(R(\mathbf{W},\boldsymbol{\lambda}))$.

Consider the third part of (46), $\frac{1}{n}tr(\mathbf{A}^T\mathbf{D}^{(2)}\mathbf{A}\mathbf{\Sigma})=o(1)\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma})=o(R(\mathbf{W},\boldsymbol{\lambda}))$ since all eigenvalues of $\mathbf{D}_{ii}^{(2)}+\mathbf{D}_{ii}^{(2)T}$ are of the order $o(1)$ as is shown in (45). Hence, (46) gives

$$\frac{1}{n}tr\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\Sigma\}=o(R(\mathbf{W},\boldsymbol{\lambda})).$$

Therefore, (41) has been proved.

Next, we proceed to prove (42). Define envelop matrices $\mathbf{D}^{(1)*}=\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)^*}=diag\{\mathbf{D}_{11}^{(2)^*},\ldots,\mathbf{D}_{nn}^{(2)^*}\}$, where $\mathbf{D}_{ii}^{(2)^*}=\frac{1}{2}(\mathbf{W}_i^{1/2}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{1/2}/\alpha_i+\alpha_i\mathbf{W}_i^{-1/2}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{-1/2})$ with $\alpha_i=\lambda_{max}(\mathbf{W}_i)$. It is easy to check that $\mathbf{D}^{(1)*}$ and $\mathbf{D}^{(2)*}$ are valid envelops of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, respectively. Since under Condition 5, we have

$$tr(\mathbf{D}_{ii}^{(1)}\mathbf{W}_i\mathbf{D}_{ii}^{(1)T})\leq\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})\lambda_{max}^2(\tilde{\mathbf{D}}_{ii}^{(1)})tr(\tilde{\mathbf{D}}_{ii}^{(1)2})$$
$$=\{\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2)\}\lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$$
$$=\lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2),$$
$$tr(\mathbf{D}_{ii}^{(1)*}\mathbf{W}_i)\leq\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{D}}_{ii}^{(1)2})=\lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2),$$

and

$$tr(\mathbf{D}_{ii}^{(2)}\mathbf{W}_i\mathbf{D}_{ii}^{(2)T})\leq\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{D}}_{ii}^{(2)2})=\lambda_{max}(\mathbf{W})O(n^{-4}tr(\mathbf{A})^4)$$
$$=\lambda_{max}(\mathbf{W})o(n^{-2}tr(\mathbf{A})^2),$$
$$tr(\mathbf{D}_{ii}^{(2)*}\mathbf{W}_i)\leq\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{D}}_{ii}^{(2)})=\lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2).$$

By applying Lemma A.5, we have

$$\frac{1}{n^2}Var\{\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(m)}(\mathbf{I}-\mathbf{A})\mathbf{Y}\}=o_p(R^2(\mathbf{W},\boldsymbol{\lambda})),\ m=1,2,$$

and (42) follows by the Cauchy–Schwarz inequality. □

REFERENCES

ANDERSON, T. W. and GUPTA, S. D. (1963). Some inequalities on characteristic roots of matrices. *Biometrika*, **50** 522–524.

BENASSENI, J. (2002). A complementary proof of an eigenvalue property in correspondence analysis. *Linear Algebra and Its Applications*, **354** 49–51.

CAI, T. T. and YUAN, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *Annals of Statistics*, **39** 2330–2355.

CHIANG, C. T., RICE, J. A. and WU, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, **96** 605–619.

CLAESKENS, G., KRIVOBOKOVA, T. and OPSOMER, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, **96** 529–544.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, **31** 377–403.

DIGGLE, P. J., LIANG, K. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data, 2nd Edition*. Oxford University Press, Oxford.

FAN, J. and ZHANG, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Ser. B*, **62** 303–322.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

GU, C. and MA, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *Annals of Statistics*, **33** 1357–1379.

GU, C. and WAHBA, G. (1991). Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computation*, **12** 383–398.

HAN, C. and GU, C. (2008). Optimal smoothing with correlated data. *Sankhya: The Indian Journal of Statistics*, **70** 38–72.

HOOVER, D. R., RICE, J. A., WU, C. O. and YANG, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85** 809–822.

HUANG, J. Z., WU, C. O. and ZHOU, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89** 111–128.

KASLOW, R. A., OSTROW, D. G., DETELS, R., PHAIR, J. P., POLK, B. F. and RINALDO, C. R. (1987). The multicenter aids cohort study: Rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, **126** 310–318.

LI, K. C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics*, **14** 1101–1112.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13–22.

LIN, D. Y. and YING, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*, **96** 103–126.

LIN, X. and CARROLL, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, **95** 520–534.

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Ser. B*, **53** 233–243.

WANG, L., LI, H. and HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103** 1556–1569.

WANG, N. (2003). Marginal nonparametric kernel regression accounting for within subject correlation. *Biometrika*, **90** 43–52.

WANG, N., CARROLL, R. J. and LIN, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, **100** 147–157.

WANG, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of The Royal Statistical Society Series B-statistical Methodology*, **60** 159–174.

WELSH, A. H., LIN, X. and CARROLL, R. J. (2002). Marginal longitudinal nonparametric regression. *Journal of the American Statistical Association*, **97** 482–493.

WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99** 673–686.

WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R*. CRC/Chapman and Hall, Boca Raton, Florida.

WU, C. O. and CHIANG, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, **10** 433–456.

WU, H. and ZHANG, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis*. John Wiley and Sons, Hoboken, New Jersey.

ZEGER, S. L. and 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. and SOWERS, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93** 710–719.

ZHU, Z., FUNG, W. and HE, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika*, **95** 907–917.

SUPPLEMENTARY MATERIAL

Supplement A: Efficient algorithm and additional proofs (http://lib.stat.cmu.edu/aoas/???/???). In the Supplementary Material, we give a detailed description of the algorithm proposed in Section 3.2. In addition, proofs of some technical Lemmas are also included.

GANGGANG XU
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TX 77843-3143
EMAIL: E-MAIL: gang@stat.tamu.edu

JIANHUA Z. HUANG
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TX 77843-3143
EMAIL: E-MAIL: jianhua@stat.tamu.edu