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

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

By Ganggang Xu and Jianhua Z. Huang*

Texas A&M University

This document gives the details of an efficient algorithm for optimizing the leave-subject-out crossvalidation criterion and provides additional technical proofs that can not fit in the main paper due to limit of space.

1. Efficient algorithm. We develop an efficient algorithm 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. Our algorithm can be viewed as an extension of the stable and fast algorithm of Wood (2004) to the longitudinal data case. Define the transformed data as $\tilde{\boldsymbol{Y}} = \mathbf{W}^{-1/2}\boldsymbol{Y}$, $\tilde{\mathbf{X}} = \mathbf{W}^{-1/2}\mathbf{X}$, and the corresponding hat matrix as

$$\tilde{\mathbf{A}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \boldsymbol{S})^{-1}\tilde{\mathbf{X}}^T,$$

where $\mathbf{S} = \sum_{k=1}^{m} \lambda_k \mathbf{S}_k$. Since $\mathbf{S}$ is positive semi-definite, we can find a matrix $\mathbf{B}$ with full column rank such that $\mathbf{S} = \mathbf{B}^T\mathbf{B}$ using, for example, the Cholesky decomposition. Form the QR decomposition $\tilde{\mathbf{X}} = \mathbf{Q}^T\mathbf{R}$, where $\mathbf{Q}$ is a $N \times p$ column orthonormal matrix and $\mathbf{R}$ is a $p \times p$ upper triangular matrix, N is the total number of observations in all subjects and p is the number of columns in the design matrix $\mathbf{X}$. The identity $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \mathbf{S} = \mathbf{R}^T\mathbf{R} + \mathbf{B}^T\mathbf{B}$ motivates us to form the singular value decomposition

$$\begin{pmatrix} \mathbf{R} \\ \mathbf{B} \end{pmatrix} = \mathbf{U}\mathbf{D}\mathbf{V}^T \approx \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}, \tag{S.1}$$

where $\mathbf{D}$ is the diagonal matrix of singular values, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. Some of the diagonal elements of $\mathbf{D}$ can be very small and thus can be removed without causing appreciable errors. The matrices $\mathbf{U}^*, \mathbf{D}^*, \mathbf{V}^*$ in (S.1) are obtained by removing small singular values from $\mathbf{D}$ along with

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

the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$. Define the sub matrix $\mathbf{U}_1^*$ of $\mathbf{U}^*$ such that $\mathbf{R} = \mathbf{U}_1^*\mathbf{D}^*\mathbf{V}^{*T}$, we can rewrite the matrix $\tilde{\mathbf{A}}$ as

$$\tilde{\mathbf{A}} = \mathbf{Q}^T\mathbf{R}(\mathbf{R}^T\mathbf{R} + \mathbf{B}^T\mathbf{B})^{-1}\mathbf{R}^{-1}\mathbf{Q} = \mathbf{Q}\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}^T.$$

Note that $\mathbf{Q}$ is a $N \times p$ matrix, $\mathbf{U}_1^*$ is a $p \times p$ matrix. The fast algorithm for GCV optimization in Wood (2004) makes use of the fact that $tr(\tilde{\mathbf{A}}) = tr(\mathbf{U}_1^*\mathbf{U}_1^{*T})$, which only takes $O(p^3)$ floating operations to evaluate. However, this appealing property can not be used for the evaluation of LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$. Let LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda}) = \alpha + \beta$ with

$$\alpha = \frac{1}{n}\tilde{\boldsymbol{Y}}^T(\mathbf{I} - \tilde{\mathbf{A}})\mathbf{W}(\mathbf{I} - \tilde{\mathbf{A}})\tilde{\boldsymbol{Y}},$$

$$\beta = \frac{2}{n}\tilde{\boldsymbol{Y}}^T(\mathbf{I} - \tilde{\mathbf{A}})\mathbf{W}\Big(\sum_{i=1}^{n}\mathbf{L}_i^T\mathbf{L}_i\tilde{\mathbf{A}}\mathbf{L}_i^T\mathbf{L}_i\Big)(\mathbf{I} - \tilde{\mathbf{A}})\tilde{\boldsymbol{Y}},$$

where $\mathbf{L}_i = [\mathbf{0}, \cdots, \mathbf{I}_{ii}, \cdots, \mathbf{0}]_{n_i \times N}$. To make good use of the QR decomposition given above, we define the $p \times 1$ vectors $\tilde{\boldsymbol{Y}}_{\mathbf{Q}} = \mathbf{Q}^T\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{Y}}_W = \mathbf{Q}^T\mathbf{W}\tilde{\boldsymbol{Y}}$, the $p \times p$ matrix $\mathbf{Q}_W = \mathbf{Q}^T\mathbf{W}\mathbf{Q}$ and the $n_i \times p$ matrices $\mathbf{Q}_i = \mathbf{L}_i\mathbf{Q}$, $(i = 1, \cdots, n)$. Then, α and β can be computed using

$$\alpha = \frac{1}{n}(\tilde{\boldsymbol{Y}}^T\mathbf{W}\tilde{\boldsymbol{Y}} - 2\tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_W + \tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_W\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q),$$

$$\beta = \frac{2}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q)^T\mathbf{W}_i\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q).$$

Following Gu and Wahba (1991), we define $\eta_j = log(\lambda_j)$, $j = 1, \cdots, m$, and compute the gradients and Hessian matrix of LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$ with respect to η_j's. Define $\mathbf{M}_k = \mathbf{D}^{*-1}\mathbf{V}^{*T}\mathbf{S}_k\mathbf{V}^*\mathbf{D}^{*-1}$, $\mathbf{M}_k^* = \mathbf{U}_1^*\mathbf{M}_k\mathbf{U}_1^{*T}$ and $\mathbf{K}_k = \mathbf{M}_k\mathbf{U}_1^{*T}\mathbf{Q}_W\mathbf{U}_1^*$, then

$$\frac{\partial\alpha}{\partial\eta_k} = \frac{2\lambda_k}{n}(\tilde{\boldsymbol{Y}}_Q^T\mathbf{M}_k^*\tilde{\boldsymbol{Y}}_W - \tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{K}_k\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q),$$

$$\frac{\partial\beta}{\partial\eta_k} = \frac{2\lambda_k}{n}\tilde{\boldsymbol{Y}}_Q^T\mathbf{M}_k^*\sum_{i=1}^{n}\mathbf{Q}_i^T(\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T\mathbf{W}_i)^\dagger(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q)$$
$$- \frac{2\lambda_k}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q)^T\mathbf{W}_i\mathbf{Q}_i\mathbf{M}_k^*\mathbf{Q}_i^T(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q),$$

where $(\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T\mathbf{W}_i)^\dagger = \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T\mathbf{W}_i + \mathbf{W}_i\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T$. To derive the second derivatives of LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$, define $\mathbf{H}_{jk} = \mathbf{U}_1^*(\mathbf{M}_k\mathbf{M}_j + \mathbf{M}_j\mathbf{M}_k)\mathbf{U}_1^{*T}$, and $\mathbf{G}_{jk} = \mathbf{M}_k\mathbf{K}_j + \mathbf{M}_j\mathbf{K}_k + \mathbf{M}_k\mathbf{Q}_W\mathbf{M}_j$. Then

$$\frac{\partial^2\alpha}{\partial\eta_k\partial\eta_j} = \frac{2\lambda_k\lambda_j}{n}\{\tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{G}_{jk}\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q - \tilde{\boldsymbol{Y}}_Q^T\mathbf{H}_{jk}\tilde{\boldsymbol{Y}}_W\} + \delta_k^j\frac{\partial\alpha}{\partial\eta_k},$$

$$\frac{\partial^2 \beta}{\partial \eta_k \partial \eta_j} = \mathbf{T}_{1,kj} + \mathbf{T}_{2,kj} + (\mathbf{T}_{3,kj} + \mathbf{T}_{3,jk}) + \mathbf{T}_{4,kj} + \delta_k^j \frac{\partial \beta}{\partial \eta_k},$$

where $\delta_k^j = 1$ if $k = j$ and 0 otherwise, and

$$\mathbf{T}_{1,kj} = -\frac{2\lambda_k \lambda_j}{n} \tilde{\boldsymbol{Y}}_Q^T \mathbf{H}_{kj} \sum_{i=1}^{n} \mathbf{Q}_i^T (\mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \mathbf{Q}_i^T \mathbf{W}_i)^\dagger (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q),$$

$$\mathbf{T}_{2,kj} = \frac{2\lambda_k \lambda_j}{n} \sum_{i=1}^{n} (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q)^T \mathbf{W}_i \mathbf{Q}_i \mathbf{H}_{kj} \mathbf{Q}_i^T (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q),$$

$$\mathbf{T}_{3,kj} = -\frac{2\lambda_k \lambda_j}{n} \tilde{\boldsymbol{Y}}_Q^T \mathbf{M}_k^* \sum_{i=1}^{n} \mathbf{Q}_i^T (\mathbf{W}_i \mathbf{Q}_i \mathbf{M}_j^* \mathbf{Q}_i^T)^\dagger (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q),$$

$$\mathbf{T}_{4,kj} = \frac{2\lambda_k \lambda_j}{n} \tilde{\boldsymbol{Y}}_Q^T \mathbf{M}_k^* \sum_{i=1}^{n} \mathbf{Q}_i^T (\mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \mathbf{Q}_i^T \mathbf{W}_i)^\dagger \mathbf{Q}_i \mathbf{M}_j^* \tilde{\boldsymbol{Y}}_Q.$$

Using the derived gradients and the Hessian matrix, the minimization of $\mathrm{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ can be done using the Newton–Raphson method. The key of the algorithm is the QR decomposition of $\tilde{\mathbf{X}}$ used in (S.1), which is the computationally most expensive step of the algorithm with the cost of Np^2 floating point operations. However, this QR decomposition needs only to be carried out once for all iterations of the Newton–Raphson algorithm since $\tilde{\mathbf{X}}$ does not depend on $\boldsymbol{\lambda}$. After the $\tilde{\boldsymbol{Y}}_Q$ and $\mathbf{Q}_i$'s are obtained, the evaluations of α and β cost $O(p^2)$ and $O(p^2+Np)$ floating point operations, respectively. The computation of gradients and the Hessian matrix of $\mathrm{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})$ can be efficiently computed in a similar manner as α and β by using above formulas. As a comparison, using the Newton–Raphson method to find the minimizer of the initial leave-subject-out crossvalidation criterion $\mathrm{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})$ is much more expensive. For each iteration, it involves formation of the hat matrix $\mathbf{A}$ ($O(Np^2)$ operations), the inversion of $\mathbf{A}_{ii}$'s ($O(\sum_{i=1}^n n_i^3)$ operations), and the summation ($O(\sum_{i=1}^n n_i^2)$ operations). The overall computational cost for each iteration is $O(Np^2)$, which is much more than the cost of minimizing $\mathrm{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})$ ($O(Np)$ operations) when p is large.

Our implementation of the algorithm followed suggestions of Wood (2004) on convergence criteria and choosing searching directions in each iteration. Similar reasoning as in Wood (2004) suggests that our algorithm can also be applied to each step of the iterative reweighted penalized least squares algorithm when there is a non-linear link function and/or data are generated from the exponential family distributions.

2. Additional technical proofs.

Proof of Lemma 3.1. For a fixed $\boldsymbol{\lambda}$, let $\hat{\boldsymbol{\beta}}^{[-i]}$ be the minimizer of the penalized weighted least square (2) omitting observations from subject i. Consider the data set $\{(\boldsymbol{y}_l^*, X_l)\}, 1 \le l \le n$, where $\boldsymbol{y}_i^* = \mathbf{X}_i\hat{\beta}^{[-i]}$ and $\boldsymbol{y}_l^* = \boldsymbol{y}_l$ if $l \neq i$, $l = 1, \cdots, n$. Then, for any $\boldsymbol{\beta}$,

$$\begin{aligned}
pl(\boldsymbol{\beta}) &= \sum_{l=1}^{n}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta}) + \sum_{k=1}^{m}\lambda_k\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta} \\
&\ge \sum_{l\neq i}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta}) + \sum_{k=1}^{m}\lambda_k\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta} \\
&\ge \sum_{l\neq i}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]}) + \sum_{k=1}^{m}\lambda_k\hat{\boldsymbol{\beta}}^{[-i]T}\mathbf{S}_k\hat{\boldsymbol{\beta}}^{[-i]} \\
&= \sum_{l=1}^{n}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]}) + \sum_{k=1}^{m}\lambda_k\hat{\boldsymbol{\beta}}^{[-i]T}\mathbf{S}_k\hat{\boldsymbol{\beta}}^{[-i]}.
\end{aligned}$$

Hence, $\hat{\boldsymbol{\beta}}^{[-i]}$ is the minimizer of $pl(\boldsymbol{\beta})$ given data $\{(\boldsymbol{y}_l^*, \mathbf{X}_l)\}$, which implies

$$\mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]} = \mathbf{L}_i\mathbf{A}(\boldsymbol{\lambda})\boldsymbol{Y}^*,$$

where $\boldsymbol{Y}^* = (\boldsymbol{y}_1^{*T}, \cdots, \boldsymbol{y}_n^{*T})^T$, and $\mathbf{L}_i = [\mathbf{0}, \cdots, \mathbf{I}_{ii}, \cdots, \mathbf{0}]_{n_i\times N}$ with $\mathbf{I}_{ii}$ being the $n_i \times n_i$ identity matrix. By the definition of $\boldsymbol{Y}^*$ and using $\mathbf{A}_{ii} = \mathbf{L}_i\mathbf{A}\mathbf{L}_i^T$, we have that

$$\mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]} = \mathbf{L}_i\mathbf{A}\left\{\boldsymbol{Y} - \mathbf{L}_i^T\left(\boldsymbol{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]}\right)\right\} = \hat{\boldsymbol{y}}_i - \mathbf{A}_{ii}(\boldsymbol{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]}).$$

By some straightforward algebra, we have that

$$(\boldsymbol{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]}) = (\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i),$$

Plugging this identity into the definition of LsoCV$(\mathbf{W}, \boldsymbol{\lambda})$, we obtain

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

which is the desired formula. □

Proof of Lemma A.4. Since $E(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) = tr(\mathbf{B})$, we have that

$$Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) = E\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\boldsymbol{e}_i^T\mathbf{B}_{ij}\boldsymbol{e}_j^T\boldsymbol{e}_k^T\mathbf{B}_{lk}^T\boldsymbol{e}_l\Big) - \{tr(\mathbf{B})\}^2.$$

Using the fact that $\boldsymbol{e}_i$'s are independent and $E(\boldsymbol{e}_i) = 0$, we obtain

$$\begin{aligned} Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) &= \sum_{i=1}^{n} E(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i)^2 + \sum_{i\neq j=1}^{n} E(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i\boldsymbol{e}_j^T\mathbf{B}_{jj}^T\boldsymbol{e}_j) \\ &\quad + 2\sum_{i\neq j=1}^{n} E(\boldsymbol{e}_i^T\mathbf{B}_{ij}\boldsymbol{e}_j\boldsymbol{e}_j^T\mathbf{B}_{ij}^T\boldsymbol{e}_i) - \{tr(\mathbf{B})\}^2 \\ &= \sum_{i=1}^{n} E(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i)^2 + \sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ii})tr(\mathbf{B}_{jj}^T) \\ &\quad + 2\sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ij}\mathbf{B}_{ij}^T) - \{tr(\mathbf{B})\}^2. \end{aligned}$$

Notice that

$$tr(\mathbf{B}\mathbf{B}^T) = \sum_{i=1}^{n} tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T) + \sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ij}\mathbf{B}_{ij}^T),$$

$$\{tr(\mathbf{B})\}^2 = \sum_{i=1}^{n}\{tr(\mathbf{B}_{ii})\}^2 + \sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ii})tr(\mathbf{B}_{jj}^T),$$

$$\{E(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i)\}^2 = \{tr(\mathbf{B}_{ii})\}^2.$$

Some straightforward algebra yield

$$Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) = 2tr(\mathbf{B}\mathbf{B}^T) + \sum_{i=1}^{n} Var(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i) - 2\sum_{i=1}^{n} tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T).$$

Consider the eigen decomposition $\mathbf{B}_{ii}^* = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^T$. Let $\lambda_{ij}(\mathbf{B}_{ii}^*) \geq 0$ be the jth eigenvalue, and $\mathbf{u}_{ij}$ be the jth column of $\mathbf{U}_i$, $j = 1, \cdots, n_i$. Define $z_{ij} = (\mathbf{u}_{ij}^T\boldsymbol{e}_i)^2$, then by the condition of this lemma, we have that $E(z_{ij}z_{ik}) \leq \frac{1}{2}\{E(z_{ij}^2) + E(z_{ik}^2)\} \leq K$. By some algebra, we have

$$\begin{aligned} Var(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i) &\leq E\{\boldsymbol{e}_i^T(\mathbf{B}_{ii} + \mathbf{B}_{ii}^T)\boldsymbol{e}_i/2\}^2 \leq E(\boldsymbol{e}_i^T\mathbf{B}_{ii}^*\boldsymbol{e}_i)^2 \\ &= E\Big\{\sum_{j=1}^{n_i}\lambda_{ij}(\mathbf{B}_{ii}^*)z_{ij}\Big\}^2 = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\lambda_{ij}(\mathbf{B}_{ii}^*)\lambda_{ik}(\mathbf{B}_{ii}^*)E(z_{ij}z_{ik}) \\ &\leq K\{\sum_{j=1}^{n_i}\lambda_{ij}(\mathbf{B}_{ii}^*)\}^2 = K\{tr(\mathbf{B}_{ii}^*)\}^2. \end{aligned}$$

Therefore, we get

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

Since $tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T) \geq 0$, the desired inequality holds. □

Proof of Lemma A.5. Using the decomposition $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, we obtain

$$\begin{aligned}&\frac{1}{n^2}Var\{\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\mathbf{Y}\})\\ &\quad= \frac{1}{n^2}Var\{\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon} + 2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\}.\end{aligned}$$

By a simple application of the Cauchy–Schwarz inequality, it suffices to show

$$\frac{1}{n^2}Var\{\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{A})\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon})\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})), \quad \text{and} \tag{S.2}$$

$$\frac{1}{n^2}Var\{2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{S.3}$$

We shall first show (S.2). Using Lemma A.4 with $\boldsymbol{e} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$, $\mathbf{B} = \boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}$ and $\mathbf{B}^* = \boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^*(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}$, one has

$$\frac{1}{n^2}Var\{\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\} \leq \frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) + \frac{K}{n^2}\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^*)\}^2. \tag{S.4}$$

Repeatedly using Lemmas A.2 and A.3, and the fact that $\lambda_{max}((\mathbf{I}-\tilde{\mathbf{A}})^2) \leq 1$, we have

$$\begin{aligned}tr(\mathbf{B}\mathbf{B}^T) &= tr\{\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\}\\ &\leq \lambda_{max}\{(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}(\mathbf{I}-\tilde{\mathbf{A}})\}\\ &\qquad\qquad \times \lambda_{max}\{(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}\}tr(\mathbf{D}\mathbf{W}\mathbf{D}^T)\\ &\leq \lambda_{max}^2(\boldsymbol{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})tr(\mathbf{D}\mathbf{W}\mathbf{D}^T).\end{aligned}$$

Noticing that $\max_{1\leq i\leq n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii}^T)\} = \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$ and using Conditions 3–4, we have

$$\frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) = \frac{2\xi^2(\boldsymbol{\Sigma},\mathbf{W})}{n}O(n^{-2}tr(\mathbf{A})^2) = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{S.5}$$

Note that $\mathbf{B}_{ii}^* = \mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^*(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T$, where $\mathbf{L}_i$ is a selection

matrix as used in the proof of Lemma 3.1. Thus, using lemma A.2,

$$\begin{aligned} tr(\mathbf{B}_{ii}^*) &= tr\{\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^*(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\} \\ &= tr(\boldsymbol{\Sigma}_i\mathbf{D}_{ii}^*) - tr\{\boldsymbol{\Sigma}_i(\mathbf{A}_{ii}^T\mathbf{D}_{ii}^* + \mathbf{D}_{ii}^*\mathbf{A}_{ii})\} \\ &\qquad + tr\{\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{A}^T\mathbf{D}^*\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\} \\ &\le tr(\boldsymbol{\Sigma}_i\mathbf{D}_{ii}^*) - tr\{\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1/2}(\tilde{\mathbf{A}}_{ii}\tilde{\mathbf{D}}_{ii}^* + \tilde{\mathbf{D}}_{ii}^*\tilde{\mathbf{A}}_{ii})\} \\ &\qquad + \lambda_{max}(\mathbf{W}^{1/2}\mathbf{D}^*\mathbf{W}^{1/2})tr(\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1/2}\mathbf{C}_{ii}) \\ &\le \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})tr(\mathbf{D}_{ii}^*\mathbf{W}_i) - tr\{\tilde{\boldsymbol{\Sigma}}_i(\tilde{\mathbf{A}}_{ii}\tilde{\mathbf{D}}_{ii}^* + \tilde{\mathbf{D}}_{ii}^*\tilde{\mathbf{A}}_{ii})\} \\ &\qquad + \max_{1\le i\le n}\{tr(\mathbf{D}_{ii}^*\mathbf{W}_i)\}\lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})tr(\mathbf{C}_{ii}), \end{aligned} \tag{S.6}$$

where $\mathbf{C}_{ii}$ is the ith diagonal block of $\tilde{\mathbf{A}}^2$, $\tilde{\boldsymbol{\Sigma}}_i = \mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1/2}$ and $\tilde{\mathbf{D}}_{ii}^* = \mathbf{W}_i^{1/2}\mathbf{D}_{ii}^*\mathbf{W}_i^{1/2}$. By Condition 2, $tr(\mathbf{C}_{ii}) = o(1)$ and thus the third term in (S.6) can be ignored compared to the first term, which is of the order $\xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-1}tr(\mathbf{A}))$ by the condition (i) in this Lemma. Now consider the second term in (S.6). Since $tr\{\tilde{\boldsymbol{\Sigma}}_i(\alpha\tilde{\mathbf{A}}_{ii} + \tilde{\mathbf{D}}_{ii}^*)^2\} \ge 0$ for any α, one has that

$$\begin{aligned} &- tr\{\boldsymbol{\Sigma}_i(\mathbf{A}_{ii}^T\mathbf{D}_{ii}^* + \mathbf{D}_{ii}^*\mathbf{A}_{ii})\} \\ &\quad \le \alpha tr(\tilde{\boldsymbol{\Sigma}}_i\tilde{\mathbf{A}}_{ii}^2) + tr(\tilde{\boldsymbol{\Sigma}}_i\tilde{\mathbf{D}}_{ii}^{*2})/\alpha \\ &\quad \le \lambda_{max}(\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})tr(\alpha\tilde{\mathbf{A}}_{ii}^2 + \tilde{\mathbf{D}}_{ii}^{*2}/\alpha) \\ &\quad \le \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})[\alpha\{tr(\tilde{\mathbf{A}}_{ii})\}^2 + \{tr(\mathbf{D}_{ii}\mathbf{W}_i)\}^2/\alpha] \\ &\quad = \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2). \end{aligned}$$

The last equation is obtained by taking $\alpha = \lambda_{max}(\mathbf{W})$. Recall that $O(tr(\mathbf{A})/n) = o(1)$, by equation (S.6), using Conditions 3–4, we have $tr(\mathbf{B}_{ii}^*) \sim \xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-1}tr(\mathbf{A}))$ and thus

$$\frac{K}{n^2}\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii}^*)\}^2 = \frac{1}{n}\xi^2(\boldsymbol{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{S.7}$$

Combining equations (S.5)-(S.7), we obtain (S.2).

To show (S.3), note that

$$\lambda_{max}(\mathbf{D}\mathbf{W}\mathbf{D}^T) \le \max_{1\le i\le n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii}^T)\} = \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2).$$

Use Lemma A.3 and Conditions 3–4 to yield

$$\begin{aligned}
&\frac{1}{n^2}Var\{2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\} \\
&\quad= \frac{4}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad\le \frac{4\lambda_{max}\{(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}(\mathbf{I}-\tilde{\mathbf{A}})\}}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}\mathbf{W}\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad\le \frac{4\lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})}{n}\lambda_{max}(\mathbf{D}\mathbf{W}\mathbf{D}^T)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad= \frac{\xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad= o(R^2(\mathbf{W},\boldsymbol{\lambda})),
\end{aligned}$$

which is the desired result. □

References.

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.

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

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