Optimal Smoothing in Nonparametric Mixed-Effect Models

Chong Gu* and Ping Ma†
Purdue University and Harvard University

Abstract

Mixed-effect models are widely used for the analysis of correlated data such as longitudinal data and repeated measures. In this article, we study an approach to the nonparametric estimation of mixed-effect models. We consider models with parametric random effects and flexible fixed effects, and employ the penalized least squares method to estimate the models. The issue to be addressed is the selection of smoothing parameters through the generalized cross-validation method, which is shown to yield optimal smoothing for both real and latent random effects. Simulation studies are conducted to investigate the empirical performance of generalized cross-validation in the context. Real data examples are presented to demonstrate the applications of the methodology.

AMS 2000 subject classifications. Primary 62G08; secondary 62G05, 62G20, 62H12, 41A15.

Key words and phrases. Correlated error, generalized cross-validation, longitudinal data, mixed-effect model, penalized least squares, repeated measure, smoothing spline.

1 Introduction

Mixed-effect models are widely used for the analysis of data with correlated errors. The linear mixed-effect models, also known as variance component models, are of the form

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{b} + \epsilon_i, \tag{1.1}$$

$i = 1, \ldots, n$, where $\mathbf{x}_i^T \boldsymbol{\beta}$ are the fixed effects, $\mathbf{z}_i^T \mathbf{b}$ are the random effects with $\mathbf{b} \sim N(\mathbf{0}, B)$, and $\epsilon_i \sim N(0, \sigma^2)$ are independent of $\mathbf{b}$ and of each other; see, e.g., Harville (1977) and Robinson (1991). The unknown parameters are $\boldsymbol{\beta}$, B, and σ^2, which are to be estimated from the data.

*Research supported by NIH Grant R33HL68515.
†Work done while a graduate student at Purdue University supported by NIH Grant R33HL68515.

Nonlinear and nonparametric generalizations of (1.1) can be found in, e.g., Lindstrom and Bates (1990), Wang (1998a), and Ke and Wang (2001).

In this article, we consider models of the form

$$Y_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b} + \epsilon_i, \tag{1.2}$$

where the regression function $\eta(x)$ is assumed to be a smooth function on a generic domain $\mathcal{X}$. The model terms $\eta(x)$ or $\eta(x) + \mathbf{z}^T\mathbf{b}$ will be estimated using the penalized (unweighted) least squares method through the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta(x_i) - \mathbf{z}_i^T\mathbf{b})^2 + \frac{1}{n}\mathbf{b}^T\Sigma\mathbf{b} + \lambda J(\eta), \tag{1.3}$$

where the quadratic functional $J(\eta)$ quantifies the roughness of η and the smoothing parameter λ controls the trade-off between the goodness-of-fit and the smoothness of η; note that if one substitutes $\sigma^2 B^{-1}$ for Σ in (1.3) then the first two terms are proportional to the minus log likelihood of $(\mathbf{Y}, \mathbf{b})$. We will treat Σ as a tuning parameter like λ, however, and not be concerned with the estimation of $\sigma^2 B^{-1}$. Technically, (1.3) resembles the partial spline models, but with the partial terms $\mathbf{z}^T\mathbf{b}$ penalized.

Absent of the random effects $\mathbf{z}^T\mathbf{b}$, penalized least squares regression has been studied extensively in the literature; see, e.g., Wahba (1990) and Gu (2002) for comprehensive treatments of the subject. The models of (1.2) were first considered by Wang (1998a), who used penalized marginal likelihood (of $\mathbf{Y}$) to estimate η. Smoothing parameter selection in penalized marginal likelihood estimation with correlated data was studied by Wang (1998b), who illustrated the middling performance of various versions of cross-validation, in contrast to the more reliable performance of the generalized maximum likelihood method of Wahba (1985) derived under the Bayes model of smoothing splines. Under the Bayes model, η itself is decomposed into fixed and random effects, with $\lambda J(\eta)$ acting as the minus log likelihood of the random effects; the generalized maximum likelihood method of Wahba (1985) is essentially the popular restricted maximum likelihood method widely used for the estimation of variance component models.

The purpose of this article is to study the estimation of the model terms in (1.2) through the minimization of (1.3), with the smoothing parameter λ and the correlation parameters Σ selected by the standard generalized cross-validation method of Craven and Wahba (1979), which was developed for independent data. In some applications, the random effects $\mathbf{z}^T\mathbf{b}$ are physically interpretable, or real, and in some others, $\mathbf{z}^T\mathbf{b}$ are merely a convenient device for the modeling of variance components, or latent; for the latter case, the estimation through (1.3) turns the variance components into "mean components." For both real and latent random effects, generalized cross-validation will be shown to yield optimal smoothing, through asymptotic analysis and numerical simulation. Real-data examples are also presented to illustrate the applications of the methodology.

The rest of the article is organized as follows. In §2, the problem is formulated and preliminary analysis is conducted. Examples are given in §3. Generalized cross-validation and its optimality are discussed in §4, followed by simulation studies in §5. Real-data examples are shown in §6. Proofs of the theorems and lemmas in §4 are collected in §7. A few remarks in §8 conclude the article.

2 Penalized Least Squares Estimation

Consider the minimization of (1.3) for η in a q-dimensional space $\mathrm{span}\{\xi_1, \ldots, \xi_q\}$. Functions in the space can be expressed as

$$\eta(x) = \sum_{j=1}^{q} c_j \xi_j(x) = \boldsymbol{\xi}^T(x)\mathbf{c}. \tag{2.1}$$

Plugging (2.1) into (1.3), one minimizes

$$(\mathbf{Y} - R\mathbf{c} - Z\mathbf{b})^T(\mathbf{Y} - R\mathbf{c} - Z\mathbf{b}) + \mathbf{b}^T\Sigma\mathbf{b} + n\lambda\mathbf{c}^T Q\mathbf{c} \tag{2.2}$$

with respect to $\mathbf{c}$ and $\mathbf{b}$, where $\Sigma > 0$ is $p \times p$, R is $n \times q$ with the (i,j)th entry $\xi_j(x_i)$, $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ is $n \times p$, and Q is $q \times q$ with the $(j,k)$the entry $J(\xi_j, \xi_k)$. Differentiating (2.2) with respect to $\mathbf{c}$ and $\mathbf{b}$ and setting the derivatives to 0, one has

$$\begin{pmatrix} R^TR + n\lambda Q & R^TZ \\ Z^TR & Z^TZ + \Sigma \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} R^T\mathbf{Y} \\ Z^T\mathbf{Y} \end{pmatrix}. \tag{2.3}$$

Assume that the linear system is solvable, i.e., the columns of $\binom{R^T}{Z^T}$ are in the column space of the left-hand side matrix. A solution of (2.3) is then given by

$$\begin{pmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} R^TR + n\lambda Q & R^TZ \\ Z^TR & Z^TZ + \Sigma \end{pmatrix}^+ \begin{pmatrix} R^T\mathbf{Y} \\ Z^T\mathbf{Y} \end{pmatrix},$$

where C^+ denotes the Moore-Penrose inverse of C satisfying $CC^+C = C$, $C^+CC^+ = C^+$, $(CC^+)^T = CC^+$, and $(C^+C)^T = C^+C$.

Write $D = Z^T Z + \Sigma$ and $E = (R^T R + n\lambda Q) - R^T Z D^{-1} Z^T R$. With (2.3) solvable, one has

$$\begin{pmatrix} R^T R + n\lambda Q & R^T Z \\ Z^T R & D \end{pmatrix} \begin{pmatrix} K \\ L \end{pmatrix} = \begin{pmatrix} R^T \\ Z^T \end{pmatrix}$$

for some K and L, which, after some algebra, yields $EK(I - ZD^{-1}Z^T)^{-1} = R^T$, so the columns of R^T are in the column space of E. It follows that $EE^+R^T = R^T$, and in turn

$$\begin{pmatrix} R^T R + n\lambda Q & R^T Z \\ Z^T R & Z^T Z + \Sigma \end{pmatrix}^+ = \begin{pmatrix} E^+ & -E^+ R^T Z D^{-1} \\ -D^{-1} Z^T R E^+ & D^{-1} + D^{-1} Z^T R E^+ R^T Z D^{-1} \end{pmatrix}.$$

It then follows that

$$\hat{\boldsymbol{\eta}} = R\hat{\mathbf{c}} = RE^+R^T(I - ZD^{-1}Z^T)\mathbf{Y} = M\mathbf{Y}. \tag{2.4}$$

Similarly, one has

$$\hat{\mathbf{Y}} = R\hat{\mathbf{c}} + Z\hat{\mathbf{b}} = \{(I - ZD^{-1}Z^T)RE^+R^T(I - ZD^{-1}Z^T) + ZD^{-1}Z^T\}\mathbf{Y} = A(\lambda, \Sigma)\mathbf{Y},$$

where

$$\begin{aligned} A(\lambda, \Sigma) &= (R, Z) \begin{pmatrix} R^T R + n\lambda Q & R^T Z \\ Z^T R & Z^T Z + \Sigma \end{pmatrix}^+ \begin{pmatrix} R^T \\ Z^T \end{pmatrix} \\ &= (I - ZD^{-1}Z^T)RE^+R^T(I - ZD^{-1}Z^T) + ZD^{-1}Z^T \end{aligned} \tag{2.5}$$

is known as the smoothing matrix. Alternatively, for $\tilde{E} = R^T R + n\lambda Q$ and $\tilde{D} = D - Z^T R\tilde{E}^+ R^T Z$, one may write

$$\begin{pmatrix} R^T R + n\lambda Q & R^T Z \\ Z^T R & Z^T Z + \Sigma \end{pmatrix}^+ = \begin{pmatrix} \tilde{E}^+ + \tilde{E}^+ R^T Z \tilde{D}^{-1} Z^T R \tilde{E}^+ & -\tilde{E}^+ R^T Z \tilde{D}^{-1} \\ -\tilde{D}^{-1} Z^T R \tilde{E}^+ & \tilde{D}^{-1} \end{pmatrix},$$

yielding the expressions

$$M = \tilde{A}(\lambda) - \tilde{A}(\lambda)Z(Z^T(I - \tilde{A}(\lambda))Z + \Sigma)^{-1}Z^T(I - \tilde{A}(\lambda)), \tag{2.6}$$

where $\tilde{A}(\lambda) = R\tilde{E}^+R^T$ is the smoothing matrix when the random effects are absent, and

$$A(\lambda, \Sigma) = \tilde{A}(\lambda) + (I - \tilde{A}(\lambda))Z(Z^T(I - \tilde{A}(\lambda))Z + \Sigma)^{-1}Z^T(I - \tilde{A}(\lambda)). \tag{2.7}$$

The eigenvalues of $A(\lambda, \Sigma)$ and $\tilde{A}(\lambda)$ are in the range $[0, 1]$.

With the standard formulation of penalized least squares regression, the minimization of (1.3) is performed in a so-called reproducing kernel Hilbert space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a square semi norm, and the solution resides in the space $\mathcal{N}_J \oplus \text{span}\{R_J(x_i, \cdot), i = 1, \ldots, n\}$, where $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ is the null space of $J(\eta)$ and $R_J(\cdot, \cdot)$ is the so-called reproducing kernel in $\mathcal{H} \ominus \mathcal{N}_J$. The solution has an expression

$$\eta(x) = \sum_{i=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} \tilde{c}_i R_J(x_i, x), \tag{2.8}$$

where $\{\phi_\nu\}_{\nu=1}^m$ is a basis of $\mathcal{N}_J$. It follows that $R = (S, \tilde{Q})$, where S is $n \times m$ with the (i, ν)th entry $\phi_\nu(x_i)$ and $\tilde{Q}$ is $n \times n$ with the (i, j)th entry $R_J(x_i, x_j)$. From the property of reproducing kernel, it can also be shown that $J(R_J(x_i, \cdot), R_J(x_j, \cdot)) = R_J(x_i, x_j)$, so $Q = \text{diag}(O, \tilde{Q})$. See, e.g., Wahba (1990) and Gu (2002). The linear system (2.3) is thus solvable as long as Z is of full column rank.

For fast computation, Kim and Gu (2004) consider the space $\mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, \tilde{q}\}$, where $\{z_j\}$ are a random subset of $\{x_i\}$. In that setting, $R = (S, \tilde{R})$, where $\tilde{R}$ is $n \times \tilde{q}$ with the (i, j)th entry $R_J(z_j, x_i)$, and $Q = \text{diag}(O, \tilde{Q})$, where $\tilde{Q}$ is $\tilde{q} \times \tilde{q}$ with the (j, k)th entry $R_J(z_j, z_k)$. Since $J(\eta)$ is a square norm in $\text{span}\{R_J(z_j, \cdot), j = 1, \ldots, \tilde{q}\}$, it can be shown that the columns of $\tilde{R}^T$ are in the column space of $\tilde{Q}$. It then follows that the linear system (2.3) is solvable when Z is of full column rank.

The formulation of (2.1) and (2.2) also covers general penalized regression splines, so long as (2.3) is solvable. A sufficient condition is for both R and Z to be of full column rank.

3 Examples

A few examples are in order to illustrate the formulation of the problem and the potential applications of the method under study. The examples will be employed in the simulation study of §5 and the data analysis of §6.

Example 3.1 (Growth curves) Consider the "growth" over time of a certain quantity associated with p subjects,

$$Y_i = \eta(x_i) + b_{s_i} + \epsilon_i,$$

where Y_i is the ith observation taken at time $x_i \in [0, a]$ from subject $s_i \in \{1, \ldots, p\}$, $b_s \sim N(0, \sigma_s^2)$ is the subject random effect, independent of the measurement error ϵ_i and of each other. In this setting, $B = \sigma_s^2 I$, so the $p \times p$ matrix Σ is diagonal with only one tunable parameter. The random effects b_s are real.

Taking $J(\eta) = \int_0^a (d^2\eta/dx^2)^2 dx$, one has the cubic smoothing spline, with the ϕ_ν and R_J functions in (2.8) given by

$$\phi_1(x) = 1, \quad \phi_2(x) = x, \quad R_J(x_1, x_2) = \int_0^a (x_1 - u)_+ (x_2 - u)_+ du,$$

where $(\cdot)_+ = \max(\cdot, 0)$. See, e.g., Gu (2002, §2.3.1). The null space model has an expression $\eta(x) = \beta_0 + \beta_1 x$.

Taking $J(\eta) = \int_0^a (L_\theta \eta)^2 h_\theta dx$, where $L_\theta = (d/dx)(d/dx + \theta)$ and $h_\theta = e^{3\theta x}$ for some $\theta > 0$, one has an (negative) exponential spline. The null space model has an expression $\eta(x) = \beta_0 + \beta_1 e^{-\theta x}$. Transforming x by $\tilde{x} = (1 - e^{-\theta x})/\theta$, it can be shown that

$$\int_0^a (L_\theta \eta)^2 h_\theta dx = \int_0^{\tilde{a}} (d^2\eta/d\tilde{x}^2)^2 d\tilde{x},$$

where $\tilde{a} = (1 - e^{-\theta a})/\theta$, so one has a cubic spline in $\tilde{x}$. See, e.g., Gu (2002, Example 4.7, §4.3.4). Note that the exponential spline reduces to the cubic spline in x when $\theta = 0$.

Suppose Y is the logarithm of the measurement $\tilde{Y}$ satisfying a log-normal distribution with $\mu = \eta(x) + b_s$ and σ^2 a constant; the mean of $\tilde{Y}$ is known to be $\exp(\mu + \sigma^2/2)$. The null space model of the cubic spline characterizes an exponential growth curve for $\tilde{Y}$, and the null space model of the exponential spline corresponds to a Gompertz growth curve for $\tilde{Y}$. The splines allow departures from these parametric growth curves. □

Example 3.2 (Growth under treatment) Consider the setting of Example 3.1, but with the p subjects divided into t treatment groups. The fixed effect becomes $\eta(x, \tau)$, where $\tau \in \{1, \ldots, t\}$ denotes the treatment levels. For the identifiability of $\eta(x, \tau)$ and b_s, one needs more than one subject per treatment level. One may decompose

$$\eta(x, \tau) = \eta_\emptyset + \eta_1(x) + \eta_2(\tau) + \eta_{1,2}(x, \tau),$$

where $\eta_\emptyset$ is a constant, $\eta_1(x)$ is a function of x satisfying $\eta_1(0) = 0$, $\eta_2(\tau)$ is a function of τ satisfying $\sum_{\tau=1}^t \eta_2(\tau) = 0$, and $\eta_{1,2}(x, \tau)$ satisfies $\eta_{1,2}(0, \tau) = 0$, $\forall \tau$, and $\sum_{\tau=1}^t \eta_{1,2}(x, \tau) = 0$, $\forall x$. The term $\eta_\emptyset + \eta_1(x)$ is the "average growth" and the term $\eta_2(\tau) + \eta_{1,2}(x, \tau)$ is the "contrast growth".

For flexible models, one may use

$$J(\eta) = \theta_1^{-1} \int_0^a (d^2\eta_1/dx^2)^2 dx + \theta_{1,2}^{-1} \int_0^a \sum_{\tau=1}^t (d^2\eta_{1,2}/dx^2)^2 dx,$$

which has a null space $\mathcal{N}_J$ of dimension $2t$. A set of ϕ_ν are given by

$$\{1, x, I_{[\tau=j]} - 1/t, (I_{[\tau=j]} - 1/t)x, j = 1, \dots, t-1\},$$

and the function R_J is given by

$$R_J(x_1, \tau_1; x_2, \tau_2) = \theta_1 \int_0^a (x_1 - u)_+ (x_2 - u)_+ du + \theta_{1,2}(I_{[\tau_1=\tau_2]} - 1/t) \int_0^a (x_1 - u)_+ (x_2 - u)_+ du.$$

See, e.g., Gu (2002, §2.4.4, Problem 2.14(c)). To force an additive model $\eta(x, \tau) = \eta_\emptyset + \eta_1(x) + \eta_2(\tau)$, which yields parallel growth curves at different treatment levels, one may set $\theta_{1,2} = 0$ and remove $(I_{[\tau=j]} - 1/t)x$ from the list of ϕ_ν. One may also choose to transform x through $\tilde{x} = (1 - e^{-\theta x})/\theta$ and fit models on the $\tilde{x}$ scale. □

Example 3.3 (Clustered Observations) Consider observations from p clusters, such as in multicenter studies, $Y_i = \eta(x_i) + \tilde{\epsilon}_i$, where Y_i is taken from cluster c_i with covariate x_i. Observations from different clusters are independent, while observations from the same cluster may be correlated to various degrees. The intra-cluster correlation may be modeled via $\tilde{\epsilon}_i = b_{c_i} + \epsilon_i$, where $\mathbf{b} \sim N(0, B)$, with $B = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, independent of each other; the intra-cluster correlation in cluster c_i is given by $\sigma_i^2/(\sigma^2 + \sigma_i^2)$. In this setting, the $p \times p$ matrix Σ involves p tunable parameters on the diagonal. The random effects b_c are latent.

Note that the covariate x is generic, which can be univariate as in Example 3.1, or multivariate as in Example 3.2. □

4 Optimality of Generalized Cross-Validation

For the selection of the smoothing parameter λ (and others such as the θ in Example 3.1 and the θ_1 and $\theta_{1,2}$ in Example 3.2, if present) and the correlation parameters Σ, we propose to minimize the generalized cross-validation score

$$V(\lambda, \Sigma) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda, \Sigma))^2\mathbf{Y}}{\{n^{-1}\text{tr}(I - A(\lambda, \Sigma))\}^2}; \tag{4.1}$$

Σ may involve less than $p(p+1)/2$ tunable parameters. It will be shown in this section that the minimizers of $V(\lambda, \Sigma)$ yield optimal smoothing asymptotically, in the sense to be specified. Numerical verifications of the asymptotic analysis will be presented in the next section. Generalized cross-validation was proposed by Craven and Wahba (1979) for independent data, with the asymptotic optimality established by Li (1986) in that setting; see also Speckman (1985).

First consider the mean square error at the data points,

$$L_1(\lambda, \Sigma) = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - \eta(x_i) - \mathbf{z}_i^T \mathbf{b})^2, \tag{4.2}$$

which is a natural loss when the random effects $\mathbf{z}^T\mathbf{b}$ are real. Simple algebra yields

$$\begin{aligned} L_1(\lambda, \Sigma) &= \frac{1}{n}(A\mathbf{Y} - \boldsymbol{\eta} - Z\mathbf{b})^T(A\mathbf{Y} - \boldsymbol{\eta} - Z\mathbf{b}) \\ &= \frac{1}{n}(\boldsymbol{\eta} + Z\mathbf{b})^T(I - A)^2(\boldsymbol{\eta} + Z\mathbf{b}) - \frac{2}{n}(\boldsymbol{\eta} + Z\mathbf{b})^T(I - A)A\boldsymbol{\epsilon} + \frac{1}{n}\boldsymbol{\epsilon}^T A^2 \boldsymbol{\epsilon}, \end{aligned}$$

where $\boldsymbol{\eta} = (\eta(x_1), \ldots, \eta(x_n))^T$, $\mathbf{Y} = \boldsymbol{\eta} + Z\mathbf{b} + \boldsymbol{\epsilon}$, and the arguments (λ, Σ) are dropped from the notation of the smoothing matrix A. Taking expectation with respect to $\mathbf{b}$ and $\boldsymbol{\epsilon}$, the risk is seen to be

$$R_1(\lambda, \Sigma) = E[L_1(\lambda, \Sigma)] = \frac{1}{n}\boldsymbol{\eta}^T(I - A)^2\boldsymbol{\eta} + \frac{1}{n}\mathrm{tr}((I - A)^2 ZBZ^T) + \frac{\sigma^2}{n}\mathrm{tr}A^2. \tag{4.3}$$

Now, define

$$U(\lambda, \Sigma) = \frac{1}{n}\mathbf{Y}^T(I - A)^2\mathbf{Y} + \frac{2}{n}\sigma^2\mathrm{tr}A. \tag{4.4}$$

It follows that

$$U(\lambda, \Sigma) - L_1(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \frac{2}{n}(\boldsymbol{\eta} + Z\mathbf{b})^T(I - A)\boldsymbol{\epsilon} - \frac{2}{n}(\boldsymbol{\epsilon}^T A\boldsymbol{\epsilon} - \sigma^2\mathrm{tr}A). \tag{4.5}$$

We shall establish the optimality of $U(\lambda, \Sigma)$ under the following conditions.

Condition C.1 The eigenvalues of $\Sigma(Z^T(I - \tilde{A}(\lambda))Z + \Sigma)^{-1}\Sigma$ are bounded from above.

Condition C.1 holds for Σ with eigenvalues bounded from above, and for Σ of magnitude up to the order of $O(\sqrt{n})$ when the magnitude of $Z^T(I - \tilde{A}(\lambda))Z$ grows at a rate of $O(n)$.

Condition C.2 As $n \to \infty$, $nR_1(\lambda, \Sigma) \to \infty$.

The condition simply concedes that the parametric rate of $O(n^{-1})$ is not achievable. In the absence of random effects, for η satisfying $J(\eta) < \infty$ or more stringent smoothness conditions, it typically holds that $n^{-1}\boldsymbol{\eta}^T(I - \tilde{A}(\lambda))^2\boldsymbol{\eta} = O(\lambda^s)$ for some $s \in [1, 2]$, and $\mathrm{tr}\tilde{A}^2(\lambda) \asymp \lambda^{-1/r}$ as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$ for some $r > 1$, at least for univariate smoothing splines; see, e.g., Craven and Wahba (1979), Wahba (1985), and Gu (2002, §4.2.3). For the cubic splines of Example 3.1, $r = 4$.

Lemma 4.1 *Under Condition C.1, if* $n^{-1}\boldsymbol{\eta}^T(I - \tilde{A}(\lambda))^2\boldsymbol{\eta} = O(\lambda^s)$ *and* $\mathrm{tr}\tilde{A}^2(\lambda) = O(\lambda^{-1/r})$ *as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$, *then* $R_1(\lambda, \Sigma) = O(\lambda^s + n^{-1}\lambda^{-1/r} + n^{-1}p)$.

See §7 for the proof of the lemma. For fixed p, the random effects add little to the equation, and Condition C.2 is satisfied for $\lambda \to 0$, $n\lambda^{1/r} \to \infty$, and Σ of magnitude up to the order $O(\sqrt{n})$; the optimal $\lambda \asymp n^{-r/(sr+1)}$ is well within the domain. In fact, the restriction on Σ isn't really necessary for Condition C.2 but to assure that $R_1 \to 0$. When p grows with n, Condition C.2 clearly holds, though one may need to scale back the domain of Σ for $R_1 = o(1)$ to remain true.

Theorem 4.1 *Under Conditions C.1 and C.2, as $n \to \infty$, one has*

$$U(\lambda, \Sigma) - L_1(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L_1(\lambda, \Sigma)).$$

The proof of the theorem is given in §7. When the conditions of the theorem hold in a neighborhood of the optimal (λ, Σ), the minimizer of $U(\lambda, \Sigma)$ would deliver nearly the minimum loss.

The use of $U(\lambda, \Sigma)$ requires knowledge of σ^2, which usually is not available in practice. With an extra condition, the result also holds for $V(\lambda, \Sigma)$.

Condition C.3 As $n \to \infty$, $\{n^{-1}\mathrm{tr}A(\lambda, \Sigma)\}^2 / \{n^{-1}\mathrm{tr}A^2(\lambda, \Sigma)\} \to 0$.

In the absence of random effects, Condition C.3 generally holds in most settings of interest. In fact, it typically holds that $\mathrm{tr}\tilde{A}(\lambda) \asymp \lambda^{-1/r}$ as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$, of the same order as $\mathrm{tr}\tilde{A}^2(\lambda)$. See, e.g., Craven and Wahba (1979), Wahba (1985), Li (1986), and Gu (2002, §4.2.3).

Lemma 4.2 *If $\mathrm{tr}\tilde{A}(\lambda) = O(\lambda^{-1/r})$ and $\mathrm{tr}\tilde{A}^2(\lambda) \asymp \lambda^{-1/r}$ as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$, then Condition C.3 holds for p up to the order $O(\sqrt{n})$.*

The proof is to be found in §7.

Theorem 4.2 *Under Conditions C.1, C.2, and C.3, as $n \to \infty$, one has*

$$V(\lambda, \Sigma) - L_1(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L_1(\lambda, \Sigma)).$$

Proof: Given Theorem 4.1, the proof follows that of Theorem 3.3 in Gu (2002), page 66. □

We now turn to the case with latent random effects $\mathbf{z}^T\mathbf{b}$, for which the loss $L_1(\lambda, \Sigma)$ of (4.2) may not make much practical sense. Write $P_Z = Z(Z^TZ)^+Z^T$ and $P_Z^\perp = I - P_Z$. We consider the estimation of $P_Z^\perp\boldsymbol{\eta}$ by $P_Z^\perp\hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\eta}}$ is given in (2.4); the projection ensures the identifiability of the target function. Accounting for the error covariance $\sigma^2 I + ZBZ^T$, one may assess the estimation precision via the loss

$$\tilde{L}_2(\lambda, \Sigma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T P_Z^\perp(\sigma^2 I + ZBZ^T)^{-1}P_Z^\perp(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}).$$

Since $(\sigma^2 I + ZBZ^T)^{-1} = \sigma^{-2}(I - ZD_0^{-1}Z^T)$, where $D_0 = Z^T Z + \sigma^2 B^{-1}$, one may use

$$L_2(\lambda, \Sigma) = \sigma^2 \tilde{L}_2(\lambda, \Sigma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T P_Z^{\perp}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}), \tag{4.6}$$

which is independent of B. Write $Q_Z = ZD^{-1}Z^T$ and recall $M = RE^+R^T(I - Q_Z)$ from (2.4). Plugging $\hat{\boldsymbol{\eta}} = M(\boldsymbol{\eta} + Z\mathbf{b} + \boldsymbol{\epsilon})$ into (4.6) and taking expectation, one has the risk

$$\begin{aligned} R_2(\lambda, \Sigma) &= E[L_2(\lambda, \Sigma)] \\ &= \frac{1}{n}\big\{(\boldsymbol{\eta}^T (I - M)^T P_Z^{\perp}(I - M)\boldsymbol{\eta} + \mathrm{tr}(M^T P_Z^{\perp} MZBZ^T) + \sigma^2 \mathrm{tr}(M^T P_Z^{\perp} M)\big\}. \end{aligned} \tag{4.7}$$

From (2.5) and (2.4), one has

$$\begin{aligned} (I - A)\mathbf{Y} &= (I - Q_Z)(I - RE^+R^T(I - Q_Z))\mathbf{Y} \\ &= (P_Z^{\perp} + P_Z - Q_Z)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}} + Z\mathbf{b} + \boldsymbol{\epsilon}) \\ &= P_Z^{\perp}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) + (P_Z - Q_Z)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}} + Z\mathbf{b} + \boldsymbol{\epsilon}) + P_Z^{\perp}\boldsymbol{\epsilon} \\ &= P_Z^{\perp}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) + (P_Z - Q_Z)(\mathbf{Y} - \hat{\boldsymbol{\eta}}) + P_Z^{\perp}\boldsymbol{\epsilon}. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbf{Y}^T (I - A)^2 \mathbf{Y} = (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T P_Z^{\perp}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) + \boldsymbol{\epsilon}^T P_Z^{\perp}\boldsymbol{\epsilon} + 2(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T P_Z^{\perp}\boldsymbol{\epsilon} \\ + (\mathbf{Y} - \hat{\boldsymbol{\eta}})^T (P_Z - Q_Z)^2 (\mathbf{Y} - \hat{\boldsymbol{\eta}}), \end{aligned}$$

and hence

$$\begin{aligned} &U(\lambda, \Sigma) - L_2(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &\quad = \frac{1}{n}(\mathbf{Y} - \hat{\boldsymbol{\eta}})^T (P_Z - Q_Z)^2 (\mathbf{Y} - \hat{\boldsymbol{\eta}}) + \frac{2}{n}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T P_Z^{\perp}\boldsymbol{\epsilon} - \frac{1}{n}\boldsymbol{\epsilon}^T P_Z \boldsymbol{\epsilon} + \frac{2}{n}\sigma^2 \mathrm{tr}A. \end{aligned} \tag{4.8}$$

With an extra condition, $U(\lambda, \Sigma)$ and $V(\lambda, \Sigma)$ can be shown to track $L_2(\lambda, \Sigma)$ asymptotically.

Condition C.4 As $n \to \infty$, $R_1(\lambda, \Sigma) - R_2(\lambda, \Sigma) = o(R_1(\lambda, \Sigma))$.

Conditions C.2 and C.4 together imply that $R_1(\lambda, \Sigma) - R_2(\lambda, \Sigma) = o(R_2(\lambda, \Sigma))$ and $nR_2(\lambda, \Sigma) \to \infty$.

Subtracting (4.7) from (4.3), some algebra yields

$$\begin{aligned} R_1(\lambda,\Sigma) - R_2(\lambda,\Sigma) &= \frac{1}{n}\boldsymbol{\eta}^T(I-M)^T(P_Z-Q_Z)^2(I-M)\boldsymbol{\eta} \\ &+ \frac{1}{n}\mathrm{tr}(((P_Z-Q_Z)+(P_Z-Q_Z)RE^+R^T(P_Z-Q_Z))^2 ZBZ^T) \\ &+ \frac{\sigma^2}{n}\mathrm{tr}((Q_Z+(P_Z-Q_Z)M)^T(Q_Z+(P_Z-Q_Z)M)). \end{aligned} \tag{4.9}$$

The following lemma confirms the feasibility of Condition C.4 for fixed p.

Lemma 4.3 *For fixed p, if (i) $\boldsymbol{\eta}^T(I-A(\lambda,\Sigma))P_Z(I-A(\lambda,\Sigma))\boldsymbol{\eta} = o(\boldsymbol{\eta}^T(I-A(\lambda,\Sigma))^2\boldsymbol{\eta})$, (ii) $\Sigma < \rho_n Z^T Z$ for $\rho_n^2 = o(R_1)$, and (iii) $tr(Z^TZ)/n$ is bounded, then $R_1(\lambda,\Sigma) - R_2(\lambda,\Sigma) = o(R_1(\lambda,\Sigma))$.*

The proof of the lemma is given in §7. Condition (i) bars $(I-A)\boldsymbol{\eta}$ from being overloaded in the column space of Z, (ii) holds for Σ of magnitude up to the order of $O(\sqrt{n})$ when Z^TZ grows at a rate of $O(n)$, which is typical for fixed p. Alternatively, if $\rho_n = o(R_1)$ in (ii), which usually holds for bounded Σ, then (i) can be replaced by bounded $\boldsymbol{\eta}^T\boldsymbol{\eta}/n$; see the proof in §7.

Theorem 4.3 *Under Conditions C.1, C.2, and C.4, as $n \to \infty$, one has*

$$U(\lambda,\Sigma) - L_2(\lambda,\Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L_2(\lambda,\Sigma)).$$

If, in addition, Condition C.3 also holds, then

$$V(\lambda,\Sigma) - L_2(\lambda,\Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L_2(\lambda,\Sigma)).$$

The proof of the theorem is given in §7.

Up to this point, we have considered purely real and purely latent random effects. In practice, one could have a mixture of real and latent random effects in the same setting. Partition $Z = (Z_1, Z_2)$ and $\mathbf{b}^T = (\mathbf{b}_1^T, \mathbf{b}_2^T)$ and assume $\mathbf{b}_1$ and $\mathbf{b}_2$ are independent so B is block diagonal. Define

$$L_3(\lambda,\Sigma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} + Z_1\hat{\mathbf{b}}_1 - \boldsymbol{\eta} - Z_1\mathbf{b}_1)^T P_{Z_2}^{\perp}(\hat{\boldsymbol{\eta}} + Z_1\hat{\mathbf{b}}_1 - \boldsymbol{\eta} - Z_1\mathbf{b}_1) \tag{4.10}$$

and $R_3(\lambda,\Sigma) = E[L_3(\lambda,\Sigma)]$, where $P_{Z_2}^{\perp} = I - Z_2(Z_2^TZ_2)^+Z_2^T$; $L_3(\lambda,\Sigma)$ is a natural loss for $Z_1\mathbf{b}_1$ real and $Z_2\mathbf{b}_2$ latent. Replace $R_2(\lambda,\Sigma)$ in Condition C.4 by $R_3(\lambda,\Sigma)$.

Condition C.5 As $n \to \infty$, $R_1(\lambda,\Sigma) - R_3(\lambda,\Sigma) = o(R_1(\lambda,\Sigma))$.

A general result follows, of which the earlier theorems are special cases with nil Z_1 or nil Z_2.

Theorem 4.4 *Under Conditions C.1, C.2, and C.5, as $n \to \infty$, one has*

$$U(\lambda, \Sigma) - L_3(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L_3(\lambda, \Sigma)).$$

If, in addition, Condition C.3 also holds, then

$$V(\lambda, \Sigma) - L_3(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L_3(\lambda, \Sigma)).$$

The proof of the theorem follows straightforward modifications of the proof of Theorem 4.3 as given in §7.

5 Empirical Performance

We now present simple simulations to illustrate the practical performance of generalized cross-validation in the context.

5.1 Real Random Effects

First consider a setting with real random effects covered by Theorems 4.1 and 4.2. One hundred replicates of samples were generated according to

$$Y_i = \eta(x_i) + b_{s_i} + \epsilon_i, \quad i = 1, \dots, 100, \tag{5.1}$$

where $\eta(x) = 3\sin(2\pi x)$, x_i a random sample from $U(0,1)$, $\epsilon_i \sim N(0, 0.5^2)$, $b_s \sim N(0, 0.5^2)$, and $s_i \in \{1, \dots, 10\}$, 10 each. Cubic smoothing splines as described in Example 3.1 were calculated with (λ_u, Σ_u) minimizing $U(\lambda, \Sigma)$ of (4.4), (λ_v, Σ_v) minimizing $V(\lambda, \Sigma)$ of (4.1), and (λ_m, Σ_m) minimizing $L_1(\lambda, \Sigma)$ of (4.2).

The loss $L_1(\lambda, \Sigma)$ was recorded for the fits. For the V fit with (λ_v, Σ_v), the variance estimate through

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T(I - A(\lambda_v, \Sigma_v))^2\mathbf{Y}}{\mathrm{tr}(I - A(\lambda_v, \Sigma_v))} \tag{5.2}$$

was also recorded; the variance estimate was proposed by Wahba (1983) for independent data. The ratio σ^2/σ_s^2 as part of Σ was "estimated" through Σ_u, Σ_v, or Σ_m.

It is known that cross-validation may lead to severe undersmoothing on up to about 10% replicates. To circumvent the problem, a simple modification proved to be very effective in the

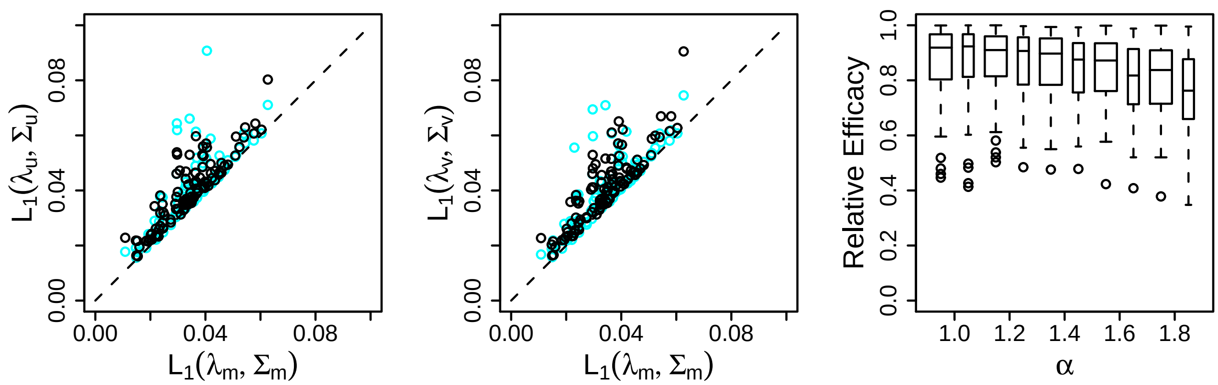

Figure 5.1: Simulation with Real Random Effects. Left and Center: Performances of $U_\alpha(\lambda, \Sigma)$ and $V_\alpha(\lambda, \Sigma)$ with $\alpha = 1$ (faded circles) and $\alpha = 1.4$ (circles). Right: $L_1(\lambda_m, \Sigma_m)/L_1(\lambda_u, \Sigma_u)$ (fatter boxes) and $L_1(\lambda_m, \Sigma_m)/L_1(\lambda_v, \Sigma_v)$ (thinner boxes) for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$.

empirical studies of Kim and Gu (2004). The modified V is given by

$$V_\alpha(\lambda, \Sigma) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda, \Sigma))^2\mathbf{Y}}{\{n^{-1}\mathrm{tr}(I - \alpha A(\lambda, \Sigma))\}^2} \tag{5.3}$$

for some $\alpha > 1$. Similarly, U can be modified by

$$U_\alpha(\lambda, \Sigma) = \frac{1}{n}\mathbf{Y}^T(I - A(\lambda, \Sigma))^2\mathbf{Y} + \frac{2}{n}\sigma^2\alpha\,\mathrm{tr}A(\lambda, \Sigma). \tag{5.4}$$

A good choice of α is around 1.4. The U and V fits with $\alpha = 1.2, 1.4, 1.6, 1.8$ were also calculated and the loss and variance estimates recorded.

The performances of $U_\alpha(\lambda, \Sigma)$ and $V_\alpha(\lambda, \Sigma)$ are illustrated in Figure 5.1. In the left and center frames, the losses $L_1(\lambda_u, \Sigma_u)$ and $L_1(\lambda_v, \Sigma_v)$ are plotted versus the minimum possible, for $\alpha = 1, 1.4$. The relative efficacy of $U_\alpha(\lambda, \Sigma)$ and $V_\alpha(\lambda, \Sigma)$ for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$ are summarized in the right frame in box plots. Roughly speaking, U_α and V_α with $\alpha = 1$ are "unbiased" by Theorems 4.1 and 4.2, and setting $\alpha > 1$ introduces "bias." The top-tier performance may degrade slightly as α increases, but the worst cases are being pulled in for α up to $1.2 \sim 1.4$, where one appears to have the "minimax" performance.

Further details of the simulation are shown in Figure 5.2. In the left frame, λ_u and λ_v for $\alpha = 1$ and $\alpha = 1.4$ are plotted against each other, where a very small λ by $\alpha = 1$ is seen to be pulled to the "normal" range by $\alpha = 1.4$. The number of cases with severe undersmoothing by cross-validation seem to be much less than what are typically seen in simulations with independent error; the phenomenon has yet to be understood. The center frame of Figure 5.2 plots the variance ratio σ^2/σ_s^2 "estimated" through Σ_m, Σ_u, and Σ_v. An interesting observation is the wide range of

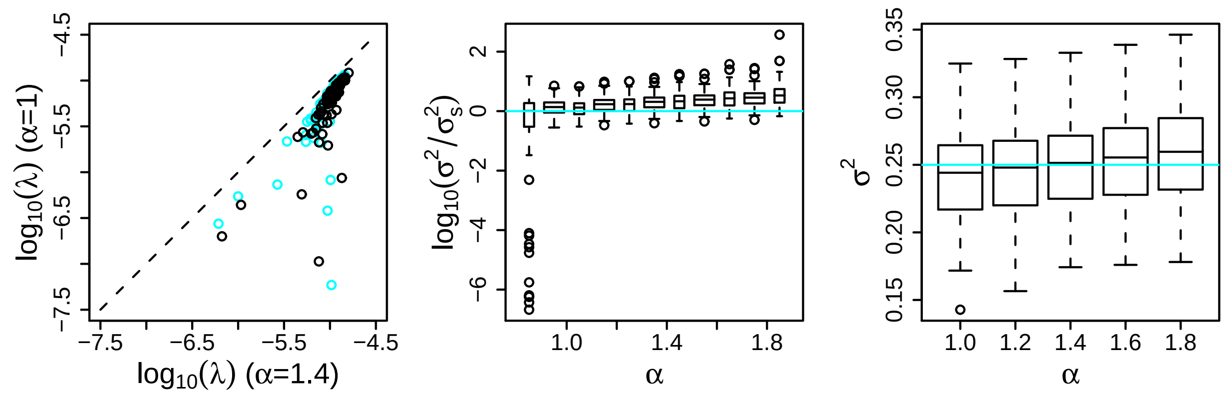

Figure 5.2: Simulation with Real Random Effects. Left: λ_u (faded circles) and λ_v (circles) for $\alpha = 1, 1.4$. Center: σ^2/σ_s^2 "Estimated" through Σ_m (left thin box), Σ_u (fatter boxes), and Σ_v (thinner boxes). Right: $\hat{\sigma}^2$. The faded horizontal lines in center and right frames mark the true values.

Σ_m, especially the many very small values, which effectively leave the term $\mathbf{z}^T\mathbf{b}$ unpenalized like the fixed effect terms in the null space of $J(\eta)$. The "estimates" through Σ_u and Σ_v appear far better in comparison, but remain highly unreliable. The upward trend of Σ_u and Σ_v with increasing α is somewhat expected, as larger α yields smoother estimates corresponding to larger penalty terms. In the right frame of Figure 5.2, the variance estimates by (5.2) are shown in box plots for V fits with $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, demonstrating generally adequate performance.

5.2 Latent Random Effects

For latent random effects, we keep the setting of (5.1) but replace b_{s_i} by b_{c_i}, as in Example 3.3. One hundred replicates of samples were generated with $\eta(x_i)$ and ϵ_i as in §5.1, and with $c_i \in \{1, 2\}$, 50 each, $b_1 \sim N(0, \sigma_1^2)$ for $\sigma_1^2 = 0.5^2$, and $b_2 \sim N(0, \sigma_2^2)$ for $\sigma_2^2 = 0.3^2$; the intra-center correlations are $0.25/(0.25 + 0.25) = 0.5$ for $c = 1$ and $0.09/(0.09 + 0.25) = 0.265$ for $c = 2$. Cubic smoothing splines were calculated with λ and Σ minimizing $U(\lambda, \Sigma)$, $V(\lambda, \Sigma)$, and $L_2(\lambda, \Sigma)$ of (4.6).

The simulation results are summarized in Figures 5.3 and 5.4. Figures 5.3 parallels Figures 5.1, except that $L_1(\lambda, \Sigma)$ is replaced by $L_2(\lambda, \Sigma)$. The left and center frames of Figures 5.4 summarize the "estimation" of the two parameters of Σ; note that the data contain only one "sample" from $N(0, \sigma_1^2)$ and one from $N(0, \sigma_2^2)$.

5.3 Mixture Random Effects

For mixture random effects, we simply add together b_s of §5.1 and b_c of §5.2, with the 10 subjects nested under the 2 clusters, 5 each. One hundred replicates of samples were generated, with the

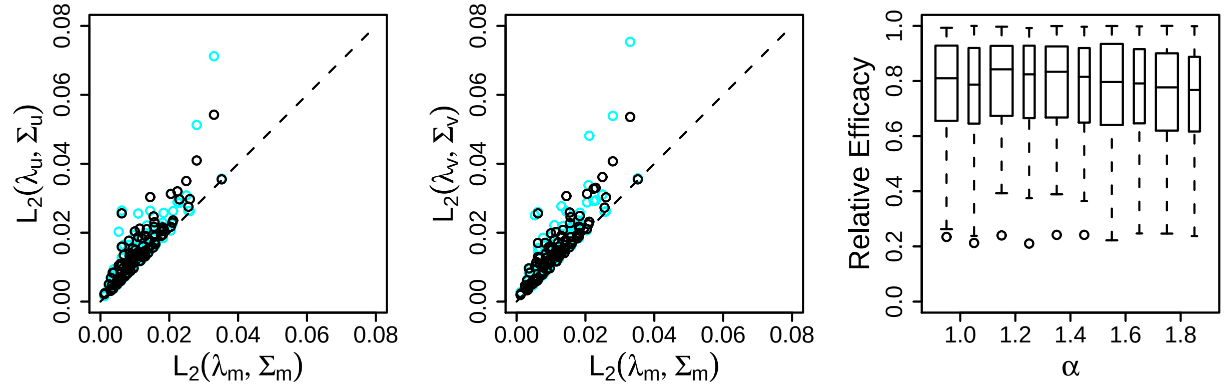

Figure 5.3: Simulation with Latent Random Effects. Left and Center: Performances of $U_\alpha(\lambda, \Sigma)$ and $V_\alpha(\lambda, \Sigma)$ with $\alpha = 1$ (faded circles) and $\alpha = 1.4$ (circles). Right: $L_2(\lambda_m, \Sigma_m)/L_2(\lambda_u, \Sigma_u)$ (fatter boxes) and $L_2(\lambda_m, \Sigma_m)/L_2(\lambda_v, \Sigma_v)$ (thinner boxes) for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$.

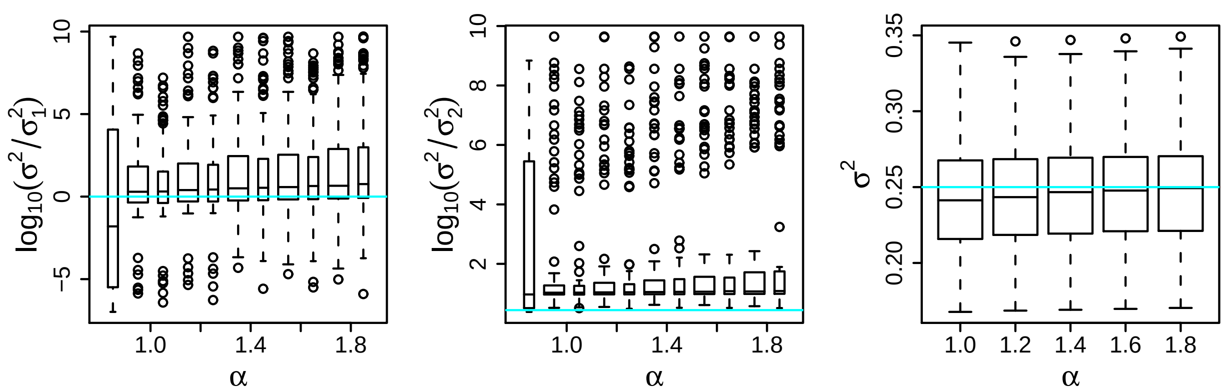

Figure 5.4: Simulation with Latent Random Effects. Left and Center: σ^2/σ_1^2 and σ^2/σ_2^2 "estimated" through Σ_m (left thin box), Σ_u (fatter boxes), and Σ_v (thinner boxes). Right: $\hat{\sigma}^2$. The faded horizontal lines mark the true values.

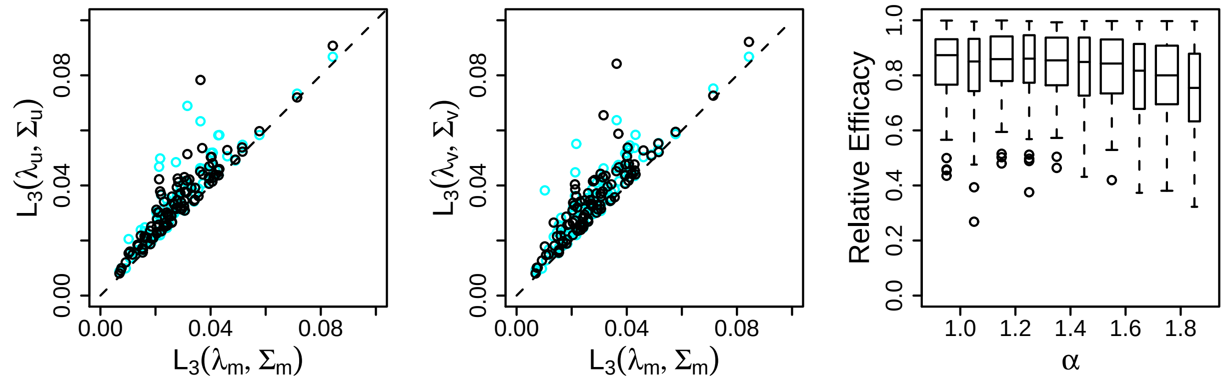

Figure 5.5: Simulation with Mixture Random Effects. Left and Center: Performances of $U_\alpha(\lambda, \Sigma)$ and $V_\alpha(\lambda, \Sigma)$ with $\alpha = 1$ (faded circles) and $\alpha = 1.4$ (circles). Right: $L_3(\lambda_m, \Sigma_m)/L_3(\lambda_u, \Sigma_u)$ (fatter boxes) and $L_3(\lambda_m, \Sigma_m)/L_3(\lambda_v, \Sigma_v)$ (thinner boxes) for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$.

specifications of $\eta(x)$, σ^2, σ_s^2, σ_1^2, and σ_2^2 remaining the same as in §5.1 and §5.2. Cubic smoothing splines were calculated with λ and Σ minimizing $U(\lambda, \Sigma)$, $V(\lambda, \Sigma)$, and $L_3(\lambda, \Sigma)$ of (4.10). The counterpart of Figures 5.1 and 5.3 is shown in Firgure 5.5. The "estimated" variance ratios are again highly unreliable, whereas $\hat{\sigma}^2$ demonstrates adequate performance, as seen in Figure 5.2 and 5.4; plots are omitted.

6 Applications

We now apply the technique to analyze a couple of real data sets.

6.1 Tumor Volume

To study the sensitivity of a human prostate tumor to androgen deprivation, a preparation of the PC82 prostate cancer cell line was implanted under the skin of 8 male nude mice. After 46 days, measurable tumors appeared on all eight mice; this day is referred to as day 0. On day 32, all mice were castrated. The tumors were measured roughly weekly over a 5 month period, resulting in 16 sets of measurements on the 8 mice. Further details concerning the data can be found in Heitjan (1991a), along with some analyses using parametric models.

We performed a nonparametric analysis of the data using the techniques developed. Taking the logarithm of the measured tumor volume as the response Y, the model of Example 3.1 was considered,

$$Y_i = \eta(x_i) + b_{s_i} + \epsilon_i,$$

where $s = 1, \ldots, 8$. The exponential spline as discussed in Example 3.1 was used to estimate

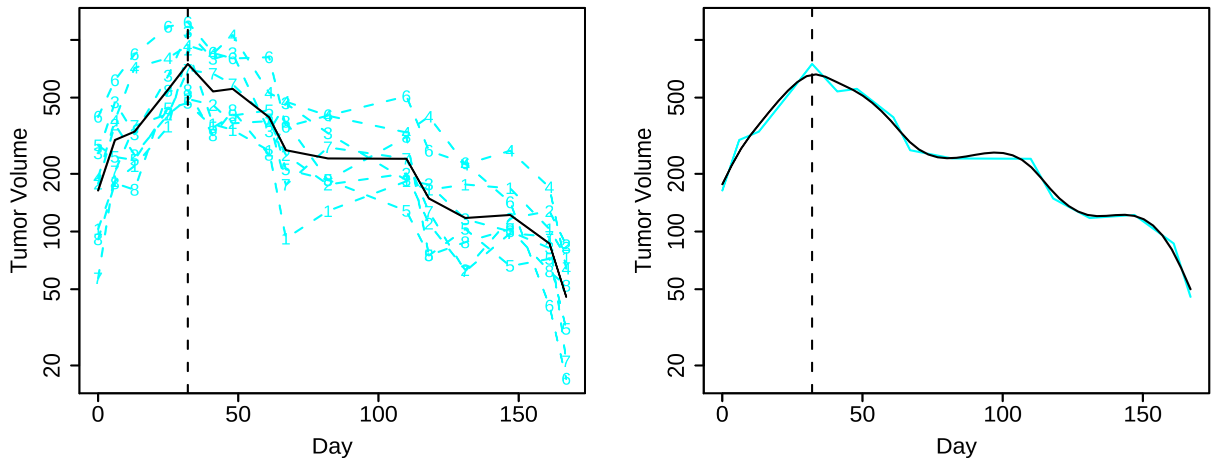

Figure 6.1: Cubic Spline Fits of Tumor Volume. Left: Tumor volume measurements (dashed lines) and their geometric mean (solid line). Right: Fitted $\eta(x)$ (solid line), with the geometric mean of measurements superimposed (faded line). The castration time is marked by the vertical line.

$\eta(x)$, but the generalized cross-validation score was minimized at $\theta = 0$, yielding a cubic spline fit. The fitted $\eta(x)$ is plotted in Figure 6.1 along with the data. The variance estimates are given by $\hat{\sigma}^2 = 0.1490$ and $\hat{\sigma}_s^2 = 0.0928$; remember that $\hat{\sigma}^2$ is trustworthy but $\hat{\sigma}_s^2$ can be grossly misleading, as shown in §5.

6.2 Treatment of Multiple Sclerosis

A randomized, double-blind clinical trial was conducted to study the treatment of multiple sclerosis by azathioprine (AZ) and methylprednisolone (MP). Patients were assigned randomly to three groups: (i) the PP group receiving placebos for both AZ and MP, (ii) the AP group receiving real AZ and placebo MP; and (iii) the AM group receiving real AZ and MP. The abundance of lymphocytes bearing a protein called F_C receptor was measured in the form of the so-called AFCR levels. Blood samples were drawn prior to the initiation of therapy, at the initiation, in weeks 4, 8, and 12, and every 12 weeks thereafter for the remainder of the trial. A total of 48 patients were represented in the data, with 17 on PP, 15 on AP, and 16 on AM. There were "missing" values in the sense that blood samples were not drawn from all patients at every time point. Detailed descriptions of the study can be found in Heitjan (1991b) and further references therein. A analysis of the data using parametric models was conducted by Heitjan (1991b).

We now present a nonparametric analysis of the data using the formulation of Example 3.2. Following Heitjan (1991b), the response Y_i are taken as the square roots of the AFCR measures.

The model is of the form

$$Y_i = \eta(x_i, \tau_i) + b_{s_i} + \epsilon_i,$$

where the patient identification s is nested under the treatment level τ. The "missing" values pose no problem for our treatment. The fitted cubic splines are plotted in Figure 6.2 with the data superimposed. The smoothing parameter $\theta_{1,2}$ was effectively set to 0 by cross-validation, so the interaction $\eta_{1,2}(x, \tau)$ consists of only parametric terms with the basis $(I_{[\tau=j]} - 1/3)x$, $j = 1, 2$; see Example 3.2 for the notation. The variance estimates were given by $\hat{\sigma}^2 = 12.81$ and $\hat{\sigma}_s^2 = 6.624$.

7 Proofs

This section collects the proofs of the lemmas and theorems of §4. The following lemmas govern some of the calculations.

Lemma 7.1 *For $Z\mathbf{b} \sim N(\mathbf{0}, ZBZ^T)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, independent of each other, one has*

$$\begin{aligned}
\text{Var}[\mathbf{b}^T Z^T C Z \mathbf{b}] &= 2\text{tr}(CZBZ^T C^T ZBZ^T), \\
\text{Var}[\boldsymbol{\epsilon}^T C \boldsymbol{\epsilon}] &= 2\sigma^4 \text{tr}(CC^T), \\
\text{Var}[\mathbf{b}^T Z^T C \boldsymbol{\epsilon}] &= \sigma^2 \text{tr}(CC^T ZBZ^T).
\end{aligned}$$

The proof of the lemma is straightforward.

Lemma 7.2 *For $M = RE^+ R^T (I - Q_Z)$, where $E = R^T (I - Q_Z) R + n\lambda Q$, one has*

$$\begin{aligned}
&M^T P_Z^\perp M + M^T (P_Z - Q_Z) M = M^T (I - Q_Z) M \le I, \\
&(I - M)^T P_Z^\perp (I - M) + (I - M)^T (P_Z - Q_Z)(I - M) = (I - M)^T (I - Q_Z)(I - M) \le 4I.
\end{aligned}$$

Proof: It is straightforward to show that $M^T (I - Q_Z) M \le I$. Now for arbitrary vector $\mathbf{x}$,

$$\mathbf{x}^T (I - M)^T (I - Q_Z)(I - M)\mathbf{x} = \mathbf{x}^T (I - Q_Z)\mathbf{x} + \mathbf{x}^T M^T (I - Q_Z) M\mathbf{x} - 2\mathbf{x}^T (I - Q_Z) M\mathbf{x}^T \le 4\mathbf{x}^T\mathbf{x},$$

where the Cauchy-Schwarz inequality is used to bound the cross term. □

Also note that B is fixed thus having bounded eigenvalues, and that $X^T X$ and XX^T share nonzero eigenvalues for all matrix X.

We are now ready for the proofs of the lemmas and theorems of §4.

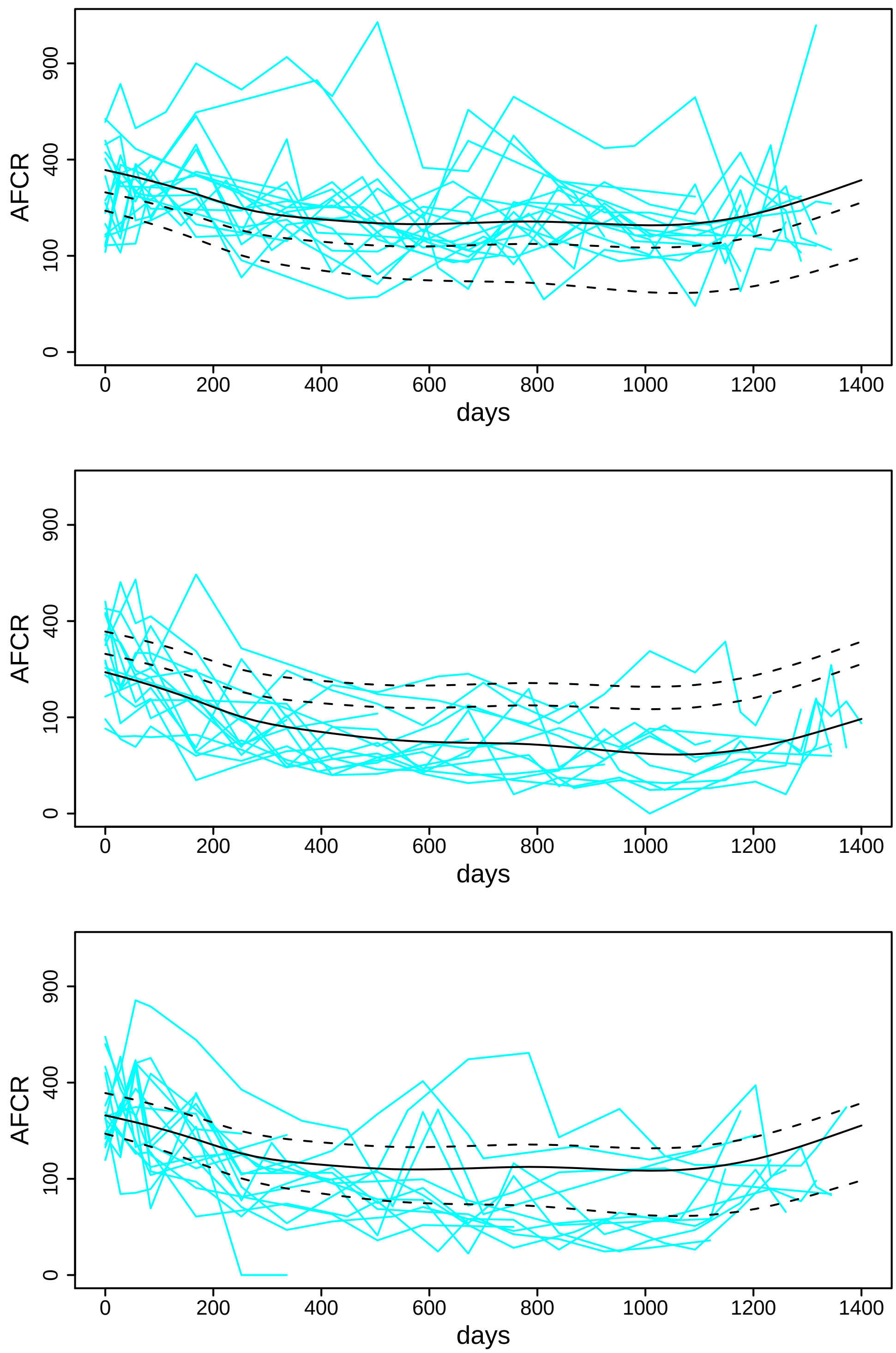

Figure 6.2: Cubic Spline Fits of AFCR Levels. From top to bottom: the PP, AP, and AM groups. The fitted $\eta(x,\tau)$, $\tau = \mathrm{PP}, \mathrm{AP}, \mathrm{AM}$, are in solid lines in their respective frames, with the corresponding data superimposed as faded lines and the other two estimates as dashed lines.

Proof of Lemma 4.1: Recall from (4.3),

$$R_1(\lambda,\Sigma) = \frac{1}{n}\boldsymbol{\eta}^T(I-A)^2\boldsymbol{\eta} + \frac{1}{n}\mathrm{tr}((I-A)^2 ZBZ^T) + \frac{\sigma^2}{n}\mathrm{tr}A^2.$$

Using (2.7), the first term is seen to be of the order $O(\lambda^s)$, and the third term is of the order $O(n^{-1}\lambda^{-1/r} + n^{-1}p)$. Again by (2.7),

$$(I-A)Z = (I-\tilde{A})Z(I-(Z^T(I-\tilde{A})Z+\Sigma)^{-1}Z^T(I-\tilde{A})Z) = (I-\tilde{A})Z(Z^T(I-\tilde{A})Z+\Sigma)^{-1}\Sigma,$$

thus

$$Z^T(I-A)^2 Z \le \Sigma(Z^T(I-\tilde{A})Z+\Sigma)^{-1}\Sigma,$$

so Condition C.1 implies an upper bound on the eigenvalues of $(I-A)ZBZ^T(I-A)$, and the second term is of the order $O(n^{-1}p)$. The proof is complete. □

Proof of Theorem 4.1: In the light of (4.5), it suffices to show that

$$L_1(\lambda,\Sigma) - R_1(\lambda,\Sigma) = o_p(R_1(\lambda,\Sigma)), \tag{7.1}$$

$$n^{-1}(\boldsymbol{\eta} + Z\mathbf{b})^T(I-A)\boldsymbol{\epsilon} = o_p(R_1(\lambda,\Sigma)), \tag{7.2}$$

$$n^{-1}(\boldsymbol{\epsilon}^T A\boldsymbol{\epsilon} - \sigma^2\mathrm{tr}A) = o_p(R_1(\lambda,\Sigma)). \tag{7.3}$$

To see (7.1), note that

$$\begin{aligned}\mathrm{Var}[L_1(\lambda,\Sigma)] = n^{-2}\mathrm{Var}[&2\boldsymbol{\eta}^T(I-A)^2 Z\mathbf{b} - 2\boldsymbol{\eta}^T(I-A)A\boldsymbol{\epsilon} \\ &+ \mathbf{b}^T Z^T(I-A)^2 Z\mathbf{b} - 2\mathbf{b}^T Z^T(I-A)A\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T A^2\boldsymbol{\epsilon}].\end{aligned}$$

Since Condition C.1 implies an upper bound on the eigenvalues of $(I-A)ZBZ^T(I-A)$, one has

$$n^{-2}\mathrm{Var}[\boldsymbol{\eta}^T(I-A)^2 Z\mathbf{b}] = n^{-2}\boldsymbol{\eta}^T(I-A)^2 ZBZ^T(I-A)^2\boldsymbol{\eta} = n^{-1}O(R_1) = o(R_1^2),$$

where the last equation is by Conditions C.2. Likewise,

$$\begin{aligned}&n^{-2}\mathrm{Var}[\boldsymbol{\eta}^T(I-A)A\boldsymbol{\epsilon}] = n^{-2}\sigma^2\boldsymbol{\eta}^T(I-A)A^2(I-A)\boldsymbol{\eta} = o(R_1^2),\\ &n^{-2}\mathrm{Var}[\mathbf{b}^T Z^T(I-A)^2 Z\mathbf{b}] = 2n^{-2}\mathrm{tr}((I-A)^2 ZBZ^T(I-A)^2 ZBZ^T) = o(R_1^2),\\ &n^{-2}\mathrm{Var}[\mathbf{b}^T Z^T(I-A)A\boldsymbol{\epsilon}] = n^{-2}\sigma^2\mathrm{tr}((I-A)A^2(I-A)ZBZ^T) = o(R_1^2),\\ &n^{-2}\mathrm{Var}[\boldsymbol{\epsilon}A^2\boldsymbol{\epsilon}] = 2n^{-2}\sigma^4\mathrm{tr}A^4 = o(R_1^2).\end{aligned}$$

Summing up, and bounding the covariances between the terms by the Cauchy-Schwarz inequality, one has $\mathrm{Var}[L_1(\lambda,\Sigma)] = o(R_1^2(\lambda,\Sigma))$, and hence (7.1). Similar calculations yield (7.2) and (7.3), completing the proof. $\square$

Proof of Lemma 4.2: From (2.7), one has $\mathrm{tr}A \le \mathrm{tr}\tilde{A} + p$ and $\mathrm{tr}A^2 \ge \mathrm{tr}\tilde{A}^2$, so

$$\frac{(n^{-1}\mathrm{tr}A)^2}{n^{-1}\mathrm{tr}A^2} \le \frac{(n^{-1}\mathrm{tr}\tilde{A} + n^{-1}p)^2}{n^{-1}\mathrm{tr}\tilde{A}^2} = O(n^{-1}\lambda^{-1/r} + n^{-1}p + n^{-1}p^2\lambda^{1/r}).$$

The lemma follows as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$. $\square$

Proof of Lemma 4.3: Recall from (4.9) that

$$\begin{aligned} R_1(\lambda,\Sigma) - R_2(\lambda,\Sigma) = & \frac{1}{n}\boldsymbol{\eta}^T(I-M)^T(P_Z-Q_Z)^2(I-M)\boldsymbol{\eta} \\ & + \frac{1}{n}\mathrm{tr}(((P_Z-Q_Z)+(P_Z-Q_Z)RE^+R^T(P_Z-Q_Z))^2 ZBZ^T) \\ & + \frac{\sigma^2}{n}\mathrm{tr}((Q_Z+(P_Z-Q_Z)M)^T(Q_Z+(P_Z-Q_Z)M)). \end{aligned}$$

Since $D = Z^TZ + \Sigma < (1+\rho_n)Z^TZ$, one has $P_Z - Q_Z < \rho_n P_Z/(1+\rho_n) < \rho_n P_Z$. For the first line, noting that $P_Z - Q_Z = P_Z(I-Q_Z)$ and $(I-Q_Z)(I-M) = I - A$, one has

$$\frac{1}{n}\boldsymbol{\eta}^T(I-M)^T(P_Z-Q_Z)^2(I-M)\boldsymbol{\eta} = \frac{1}{n}\boldsymbol{\eta}^T(I-A)P_Z(I-A)\boldsymbol{\eta} = o(R_1).$$

Alternatively, with $\boldsymbol{\eta}^T\boldsymbol{\eta}/n$ bounded,

$$\frac{1}{n}\boldsymbol{\eta}^T(I-M)^T(P_Z-Q_Z)^2(I-M)\boldsymbol{\eta} \le \rho_n\frac{1}{n}\boldsymbol{\eta}^T(I-M)^T(P_Z-Q_Z)(I-M)\boldsymbol{\eta} = O(\rho_n)$$

as $(I-M)^T(P_Z-Q_Z)(I-M) \le 4I$. For the second line, note that

$$\frac{1}{n}\mathrm{tr}((P_Z-Q_Z)ZBZ^T(P_Z-Q_Z)) \le \rho_n^2\frac{1}{n}\mathrm{tr}(ZBZ^T) = o(R_1)$$

and, with $F = (P_Z-Q_Z)^{1/2}RE^+R^T(P_Z-Q_Z)^{1/2} \le I$ and hence $F(P_Z-Q_Z)F \le \rho_n I$, that

$$\begin{aligned} & \frac{1}{n}\mathrm{tr}(((P_Z-Q_Z)RE^+R^T(P_Z-Q_Z))^2 ZBZ^T) \\ & \quad = \frac{1}{n}\mathrm{tr}(B^{1/2}Z^T(P_Z-Q_Z)^{1/2}F(P_Z-Q_Z)F(P_Z-Q_Z)^{1/2}ZB^{1/2}) \\ & \qquad \le \rho_n\frac{1}{n}\mathrm{tr}(B^{1/2}Z^T(P_Z-Q_Z)ZB^{1/2}) = \rho_n^2\frac{1}{n}\mathrm{tr}(ZBZ^T) = o(R_1); \end{aligned}$$

the cross term can be bounded by the Cauchy-Schwarz inequality. For the third line, note that $n^{-1}\mathrm{tr}Q_Z^2 \le p/n = o(R_1)$ and that $M^T(P_Z-Q_Z)^2M \le I$ has no more than p nonzero eigenvalues.

The proof is now complete. □

Proof of Theorem 4.3: Recall from (4.7) that

$$R_2(\lambda, \Sigma) = \frac{1}{n}\{(\boldsymbol{\eta}^T(I-M)^T P_Z^{\perp}(I-M)\boldsymbol{\eta} + \mathrm{tr}(M^T P_Z^{\perp} MZBZ^T) + \sigma^2 \mathrm{tr}(M^T P_Z^{\perp} M)\}.$$

Plugging $\hat{\boldsymbol{\eta}} = M(\boldsymbol{\eta} + Z\mathbf{b} + \boldsymbol{\epsilon})$ into (4.8) and grouping terms, some algebra leads to

$$\begin{aligned} U(\lambda, \Sigma) - L_2(\lambda, \Sigma) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} =& \frac{1}{n}(\boldsymbol{\eta} + Z\mathbf{b})^T(I-M)^T(P_Z - Q_Z)^2(I-M)(\boldsymbol{\eta} + Z\mathbf{b}) \\ &+ \frac{2}{n}\boldsymbol{\eta}^T(I-M)^T(P_Z - Q_Z)^2(I-M)\boldsymbol{\epsilon} + \frac{2}{n}\boldsymbol{\eta}^T(I-M)^T P_Z^{\perp}\boldsymbol{\epsilon} \\ &+ \frac{2}{n}\mathbf{b}^T Z^T(I-M)^T(P_Z - Q_Z)^2(I-M)\boldsymbol{\epsilon} - \frac{2}{n}\mathbf{b}^T Z^T M^T P_Z^{\perp}\boldsymbol{\epsilon} \\ &+ \frac{1}{n}\boldsymbol{\epsilon}^T(Q_Z + (P_Z - Q_Z)M)^T(Q_Z + (P_Z - Q_Z)M)\boldsymbol{\epsilon} \\ &- \frac{1}{n}(\boldsymbol{\epsilon}^T A\boldsymbol{\epsilon} - \sigma^2 \mathrm{tr}A). \end{aligned} \tag{7.4}$$

To prove the first part of the theorem, it suffices to show that (7.4) is of the order $o_p(R_2(\lambda, \Sigma))$ and that

$$L_2(\lambda, \Sigma) - R_2(\lambda, \Sigma) = o_p(R_2(\lambda, \Sigma)). \tag{7.5}$$

Taking expectation of the first line of (7.4), one has

$$\begin{aligned} &\frac{1}{n}E[(\boldsymbol{\eta} + Z\mathbf{b})^T(I-M)^T(P_Z - Q_Z)^2(I-M)(\boldsymbol{\eta} + Z\mathbf{b})] \\ &\qquad = \frac{1}{n}\boldsymbol{\eta}(I-M)^T(P_Z - Q_Z)^2(I-M)\boldsymbol{\eta} \\ &\qquad\quad + \frac{1}{n}\mathrm{tr}(((P_Z - Q_Z) - (P_Z - Q_Z)RE^+R^T(P_Z - Q_Z))^2 ZBZ^T) \\ &\qquad = O(R_1 - R_2) = o(R_2), \end{aligned}$$

where Condition C.4 is used. Similarly, the expectation of the fourth line of (7.4) gives

$$\begin{aligned} &\frac{1}{n}E[\boldsymbol{\epsilon}^T(Q_Z + (P_Z - Q_Z)M)^T(Q_Z + (P_Z - Q_Z)M)\boldsymbol{\epsilon}] \\ &\qquad = \frac{\sigma^2}{n}\mathrm{tr}((Q_Z + (P_Z - Q_Z)M)^T(Q_Z + (P_Z - Q_Z)M)) = O(R_1 - R_2) = o(R_2). \end{aligned}$$

For the two terms on the second line of (7.4), noting that $(I-M)^T(P_Z - Q_Z)^2(I-M) \le 4I$,

$$n^{-2}\mathrm{Var}[\boldsymbol{\eta}^T(I-M)^T(P_Z - Q_Z)^2(I-M)\boldsymbol{\epsilon}] \le 4n^{-2}\sigma^2\boldsymbol{\eta}^T(I-M)^T(P_Z - Q_Z)^2(I-M)\boldsymbol{\eta} = o(R_2^2)$$

by Conditions C.2 and C.4, and

$$n^{-2}\mathrm{Var}[\boldsymbol{\eta}^T(I-M)^T P_Z^{\perp}\boldsymbol{\epsilon}] = n^{-2}\sigma^2\boldsymbol{\eta}^T(I-M)^T P_Z^{\perp}(I-M)\boldsymbol{\eta} = n^{-1}O(R_2) = o(R_2^2).$$

Likewise, the third line terms in (7.4) give

$$\begin{aligned} & n^{-2}\mathrm{Var}[\mathbf{b}^T Z^T(I-M)^T(P_Z-Q_Z)^2(I-M)\boldsymbol{\epsilon}] \\ & \qquad \leq 2n^{-2}\sigma^2\mathrm{tr}(((P_Z-Q_Z)-(P_Z-Q_Z)RE^+R^T(P_Z-Q_Z))^2 ZBZ^T) = o(R_2^2), \end{aligned}$$

and

$$n^{-2}\mathrm{Var}[\mathbf{b}^T Z^T M^T P_Z^{\perp}\boldsymbol{\epsilon}] = 2n^{-2}\sigma^2\mathrm{tr}(M^T P_Z^{\perp} MZBZ^T) = n^{-1}O(R_2) = o(R_2^2).$$

The fifth line of (7.4) is (7.3), which is of the order $o_p(R_1) = o_p(R_2)$ by Condition C.4. To see (7.5), note that

$$\begin{aligned} \mathrm{Var}[L_2(\lambda,\Sigma)] = n^{-2}\mathrm{Var}[& 2\boldsymbol{\eta}^T(M-I)^T P_Z^{\perp} MZ\mathbf{b} + 2\boldsymbol{\eta}^T(M-I)^T P_Z^{\perp} M\boldsymbol{\epsilon} \\ & \mathbf{b}^T Z^T M^T P_Z^{\perp} MZ\mathbf{b} + \mathbf{b}^T Z^T M^T P_Z^{\perp} M\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T M^T P_Z^{\perp} M\boldsymbol{\epsilon}]. \end{aligned}$$

Using (2.6), one has $MZ = \tilde{A}Z(Z^T(I-\tilde{A})Z+\Sigma)^{-1}\Sigma$, $P_Z^{\perp}MZ = P_Z^{\perp}(\tilde{A}-I)Z(Z^T(I-\tilde{A})Z+\Sigma)^{-1}\Sigma$, thus $Z^T M^T P_Z^{\perp} MZ \leq \Sigma(Z^T(I-\tilde{A})Z+\Sigma)^{-1}\Sigma$, so Condition C.1 implies bounded eigenvalues for $P_Z^{\perp}MZBZ^T M^T P_Z^{\perp}$. It then follows that

$$\begin{aligned} & n^{-2}\mathrm{Var}[\boldsymbol{\eta}^T(M-I)^T P_Z^{\perp} MZ\mathbf{b}] = n^{-2}\boldsymbol{\eta}^T(I-M)^T P_Z^{\perp} MZBZ^T M^T P_Z^{\perp}(I-M)\boldsymbol{\eta} = o(R_2^2), \\ & n^{-2}\mathrm{Var}[\boldsymbol{\eta}^T(M-I)^T P_Z^{\perp} M\boldsymbol{\epsilon}] = n^{-2}\sigma^2\boldsymbol{\eta}^T(I-M)^T P_Z^{\perp} MM^T P_Z^{\perp}(I-M)\boldsymbol{\eta} = o(R_2^2), \\ & n^{-2}\mathrm{Var}[\mathbf{b}^T Z^T M^T P_Z^{\perp} MZ\mathbf{b}] = 2n^{-2}\mathrm{tr}(M^T P_Z^{\perp} MZBZ^T M^T P_Z^{\perp} MZBZ^T) = o(R_2^2), \\ & n^{-2}\mathrm{Var}[\mathbf{b}^T Z^T M^T P_Z^{\perp} M\boldsymbol{\epsilon}] = n^{-2}\sigma^2\mathrm{tr}(M^T P_Z^{\perp} MZBZ^T M^T P_Z^{\perp} M) = o(R_2^2), \\ & n^{-2}\mathrm{Var}[\boldsymbol{\epsilon}^T M^T P_Z^{\perp} M\boldsymbol{\epsilon}] = 2n^{-2}\sigma^4\mathrm{tr}(M^T P_Z^{\perp} MM^T P_Z^{\perp} M) = o(R_2^2), \end{aligned}$$

Collecting terms and bounding the covariances between the terms by the Cauchy-Schwarz inequality, one has $\mathrm{Var}[L_2(\lambda,\Sigma)] = o(R_2^2(\lambda,\Sigma))$, and hence (7.5). The proof of the first part of the theorem is now complete.

Given the first part of the theorem, the second part follows from the proof of Theorem 3.3 in Gu (2002), page 66. □

8 Discussion

In this article, we studied the optimal smoothing of nonparametric mixed-effect models through generalized cross-validation. The asymptotic analysis was backed by simulation studies with sample size as small as 100. Related practical issues such as variance estimation were also explored in the simulation studies. As a sequel to this work, the optimal smoothing of non-Gaussian longitudinal data has been studied in Gu and Ma (2004) on an empirical basis. The methods have been implemented in the open-source R package `gss` by the first author.

While many correlated errors can be cast as variance components with low-rank random effects, some others do not conform, which spell the limitation of the techniques developed here; an important non-conforming case is serial or spatial correlation. On the flip side, the nonparametric $\eta(x)$ can be interpreted as a realization of a Gaussian process under the Bayes model of smoothing spline, so there remains a potential identifiability problem of some sort between $\eta(x)$ and a separate serial or spatial correlation, unless the serial or spatial correlation is independent of x. Optimal smoothing for penalized likelihood estimation with serially or spatially correlated data is treated in a recent study by Han and Gu (2004).

Acknowledgments

The authors are indebted to two referees, whose comments and suggestions led to much improved presentations. The authors are also grateful to a Co-Editor and a third referee, who spotted a slopy but critical technical flaw in an earlier version of the article.

References

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

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Gu, C. and P. Ma (2004). Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection. *J. Comput. Graph. Statist. 00*, 000–000.

Han, C. and C. Gu (2004). Optimal smoothing with correlated data. Technical report, Department of Statistics, Purdue University, West Lafayette, IN.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc. 72*, 320–340 (with discussions).

Heitjan, D. F. (1991a). Generalized Norton-Simon models of tumour growth. *Statist. Med. 10*, 1075–1088.

Heitjan, D. F. (1991b). Nonlinear modeling of serial immunologic data: a case study. *J. Amer. Statist. Assoc. 86*, 891–898.

Ke, C. and Y. Wang (2001). Semiparametric nonlinear mixed-effects models and their applications. *J. Amer. Statist. Assoc. 96*, 1272–1298 (with discussions).

Kim, Y.-J. and C. Gu (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. Roy. Statist. Soc. Ser. B 66*, 337–356.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in the ridge regression with application to spline smoothing. *Ann. Statist. 14*, 1101–1112.

Lindstrom, M. J. and D. M. Bates (1990). Nonlinear mixed-effects models for repeated measures data. *Biometrics 46*, 673–687.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of the random effects. *Statist. Sci. 6*, 15–51 (with discussions).

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist. 13*, 970–983.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B 45*, 133–150.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist. 13*, 1378–1402.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

Wang, Y. (1998a). Mixed-effects smoothing spline ANOVA. *J. Roy. Statist. Soc. Ser. B 60*, 159–174.

Wang, Y. (1998b). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc. 93*, 341–348.

Department of Statistics
Purdue University
West Lafayette, IN 47907

Department of Statistics
Harvard University
Cambridge, MA 02138