Journal of Multivariate Analysis 101 (2010) 2282–2296

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

M-type smoothing spline ANOVA for correlated data

Anna Liu [a,*], Li Qin [b,c], John Staudenmayer [a]

[a] *Department of Mathematics and Statistics, University of Massachusetts, Amherst, United States*

[b] *Statistical Center for HIV/AIDS Research & Prevention, Fred Hutchinson Cancer Research Center, United States*

[c] *Department of Biostatistics, University of Washington, United States*

ARTICLE INFO

Article history:
Received 12 February 2010
Available online 12 June 2010

AMS subject classifications:
62G08
62J10
62F35

Keywords:
Correlated data
Longitudinal data
Nonparametric regression
Resistant smoothing parameter
Robust

ABSTRACT

This paper concerns outlier robust non-parametric regression with smoothing splines for data that are possibly correlated. We define a robust smoother as the minimizer of a penalized robustified log likelihood. Our estimation algorithm uses iteratively reweighted least squares to estimate the regression function. We develop two types of robust methods for joint estimation of the smoothing parameters and the correlation parameters: indirect methods and direct methods, terms borrowed from the related generalized smoothing spline literature. The indirect methods choose those parameters by conveniently approximating the distribution of the working data at each iteration as Gaussian. The direct methods estimate those parameters to minimize an estimate of the loss between the truth and the final estimated regression. Indirect methods are computationally more efficient, but our empirical studies suggest that direct methods result in more accurate estimates. Finally, the methods are applied to a data set from a macaque Simian–Human Immunodeficiency Virus (SHIV) challenge study.

1. Introduction

Smoothing spline regression is a widely used nonparametric regression technique. It smooths data with the optimal balance between the goodness-of-fit and the smoothness of the regression functions. In the most common smoothing spline model, the measure of goodness-of-fit is the sum of squared residuals, and estimation is done by penalized least squares. That measure and estimation method are sensitive to outliers, which can lead to serious under-smoothing or over-smoothing. Robust smoothing spline models have been proposed and studied by Huber [12], Cox [3], Cantoni and Ronchetti [2] and Oh et al. [18]. In these models, the goodness-of-fit measure is replaced by a robust loss function such as the commonly used Huber's loss function, and the robust spline estimate is defined as the minimizer of that loss plus a penalty to induce smoothing. These models are often termed *M*-type smoothing spline models. A deficiency is that all existing *M*-type smoothing spline models consider independent data only. It is well known that nonparametric regression techniques, including spline models, break down in the presence of correlation, with problems typically arising in the data-driven smoothing parameter selection step [19]. Robust smoothing methods specifically designed for correlated data are therefore necessary.

This paper's novel contribution is to develop *M*-type smoothing spline models for correlated data. The presence of correlation causes several non-trivial challenges. First, it is not immediately clear what makes an appropriate robust loss function. Most robust approaches for correlated data define estimates as solutions to estimating equations, which often do not have corresponding likelihood functions. Second, after a properly defined (non-quadratic) robust loss function has

* Corresponding address: 710 N. Pleasant Street, Amherst, MA 01002, United States.
E-mail address: anna@math.umass.edu (A. Liu).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.06.001

been established, the estimation algorithm needs to be determined. Third, the choice of smoothing parameter is crucial in any smoothing technique. The importance of resistance selection has been emphasized by Cantoni and Ronchetti [2], and a good selection criterion has to take into consideration both robustness and correlation. Finally, scale parameters and the correlation matrix have to be estimated as well. We assume the correlation matrix is parameterized by a few parameters. We seek to estimate the correlation parameters and the smoothing parameters simultaneously as recommended by Hart [10] and Wang [26] for correlated data.

Huggins [14] proposed a robustified parametric likelihood for repeated measurements and showed the consistency and asymptotic normality of the parameter estimates. We adapt this method to smoothing splines and use iteratively reweighted least squares (IRLS) with Fisher scoring [6] to optimize the quadraticly penalized robustified likelihood. Similar to approaches taken in generalized smoothing splines, we develop two ways to select the smoothing parameters and the covariance parameters: indirect and direct selection. The indirect selection methods obtain parameters that are currently optimal at each iteration of the iteratively reweighted least squares algorithm. Because we solve a penalized weighted least squares problem at each iteration, existing (non-robust) simultaneous selection criteria can be used at each iteration, for example, Generalized Maximum Likelihood (GML), Generalized Cross Validation (GCV) [26], and the Mallow's C_L type criterion [8]. The direct selection methods, on the other hand, optimize explicit objective functions of the smoothing parameters and covariance parameters. These objective functions estimate the discrepancy between the true and estimated regression functions. We use this method to develop three new robust direct criteria: robust GML (rGML), robust GCV (rGCV), and robust UnBiased Risk (rUBR). These are extensions of the GML, GCV, and Mallow's C_L type criteria respectively.

In addition to building a framework for robust smoothing of correlated data, this paper also contributes to M-type smoothing splines for independent data in the following ways. When the data are independent, our IRLS based algorithm coincides with the E–S algorithm by Oh et al. [18], which is motivated by the theoretical property of the M-estimator. Our indirect smoothing parameter selection methods are the same in spirit to Oh et al. [18], but our methods allow simultaneous scale estimation. Our direct methods can be used for independent data as well, and they are new in that context.

Clustered and longitudinal data are special cases of the correlated data considered in the paper. Robust modeling for clustered data and longitudinal data has been considered in [14,23,22,20,30,17,24,1,28,11,21,25]. All these papers except for the last three considered parametric regression. Among the robust semiparametric regression methods for correlated data, ours is the first semiparametric smoothing spline model.

The paper is organized as follows. In Section 2, Gaussian smoothing spline ANOVA models are reviewed. Section 3 proposes an M-type smoothing spline ANOVA for correlated data, followed by the selection methods of smoothing parameters in Section 4. Simulations are carried out in Section 5 to evaluate the proposed methodology and an application to a CD4 kinetics data set is demonstrated in Section 6. The final section contains conclusions and further remarks.

2. SS ANOVA for correlated errors

Suppose observations are generated from the following model

$$y_i = f(\mathbf{t}_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $\mathbf{t}_i = (t_1(i), \ldots, t_d(i))'$ and $t_k(i)$ belongs to an arbitrary set $\mathcal{T}_k$, $k = 1, \ldots, d$. Assume the function f has an ANOVA-like decomposition

$$f(\mathbf{t}) = \mu + \sum_k f_k(t_k) + \sum_{k<l} f_{kl}(t_k, t_l) + \cdots.$$

A reproducing kernel Hilbert Space (RKHS) $\mathcal{H}$ on $\prod_{k=1}^d \mathcal{T}_k$ can be constructed such that (1) $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_q$, where $\mathcal{H}_0 = \text{span}\{\phi_1, \ldots, \phi_M\}$ is a finite dimensional space, and $\mathcal{H}_1, \ldots, \mathcal{H}_q$ are an orthogonal RKHS, and (2) the components of the ANOVA decomposition are projections of f onto the orthogonal subspaces. Some of the decomposition components may be eliminated to achieve model parsimony.

Let $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))'$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$. Assume $\boldsymbol{\epsilon} \sim N(0, \sigma^2 W^{-1})$, the SS ANOVA estimate of f is the solution to the following [7]:

$$\min_{f \in \mathcal{H}} \left\{ (\mathbf{y} - \mathbf{f})' W (\mathbf{y} - \mathbf{f}) + \sum_{k=1}^{q} \theta_k^{-1} \|P_k f\|^2 \right\}, \tag{2}$$

where $P_j f$ is the component of f in $\mathcal{H}_j$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ are smoothing parameters.

Let $R_k(\cdot, \cdot)$ be the reproducing kernels of $\mathcal{H}_k$. Denote $\Sigma_k = \{R_k(\mathbf{t}_i, \mathbf{t}_j)\}_{i,j=1}^n$. Denote $\Sigma_{\boldsymbol{\theta}} = \theta_1 \Sigma_1 + \cdots + \theta_q \Sigma_q$ and $T = \{\phi_v(\mathbf{t}_i)\}_{i=1,\ldots,n}^{v=1,\ldots,M}$. Let the QR decomposition of T be $T = (Q_1 Q_2)(R' \mathbf{0})'$. For fixed $\boldsymbol{\theta}$ and W, the smoothing spline estimate at the design points is $\hat{\mathbf{f}} = (\hat{f}(\mathbf{t}_1), \ldots, \hat{f}(\mathbf{t}_n)') = A_W \mathbf{y}$, where

$$A_W = I - W^{-1} Q_2 (Q_2' (\Sigma_{\boldsymbol{\theta}} + W^{-1}) Q_2)^{-1} Q_2'. \tag{3}$$

Assume that W is known up to a finite number of parameters, denoted by $\boldsymbol{\tau}$. Wang [26] proposed several joint selection methods for $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$. Among the methods developed, the GML and GCV are recommended based on finite sample performance. The GML score is defined as

$$\mathrm{GML}(\boldsymbol{\theta}, \boldsymbol{\tau}) = \frac{n^{-1}\mathbf{y}'W(I - A_W)\mathbf{y}}{|W(I - A_W)|_+^{1/(n-M)}}, \tag{4}$$

where $|.|_+$ represents the product of positive eigenvalues. The GCV score is defined as

$$\mathrm{GCV}(\boldsymbol{\theta}, \boldsymbol{\tau}) = \frac{n\|W(I - A_W)\mathbf{y}\|^2}{\mathrm{tr}(W(I - A_W))^2}. \tag{5}$$

Gu and Han [8] recently developed a modified Mallows' C_L criterion and proved its asymptotic optimality as a proxy to the Kullback–Leibler distance. We will refer the criterion as U. It is defined as

$$U(\boldsymbol{\theta}, \boldsymbol{\tau}) = \log\left\{n^{-1}\mathbf{y}'W(I - A_W)^2\mathbf{y}\right\} - \frac{1}{n}\log|W| + 2\alpha\frac{\mathrm{tr}(A_W)}{n - \mathrm{tr}(A_W)} \tag{6}$$

where the recommended α value is 1.2–1.4.

3. *M*-type SS ANOVA for correlated errors

We again consider model (1), but from now on we only assume that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 W^{-1}$. Let the Cholesky decomposition of $\sigma^{-2}W$ be $\sigma^{-2}W = VV'$ and denote the ith column of V as $v_i, = 1, \ldots, n$. Assume W has a known parametric form with unknown parameters $\boldsymbol{\tau}$. For fixed $\boldsymbol{\theta}$, $\boldsymbol{\tau}$ and σ^2, we define the robustified SS ANOVA estimate of f as the minimizer of

$$\min_{f\in\mathcal{H}}\left\{\sum_{i=1}^{n}\rho(v_i'(\mathbf{y} - \mathbf{f})) + \frac{1}{2\sigma^2}\sum_{k=1}^{q}\theta_k^{-1}\|P_k f\|^2\right\} \tag{7}$$

where ρ is a suitably chosen function that downweights outlying observations and is symmetrical about zero. We assume the existence of a solution to the above minimization problem. With a convex ρ, the existence is guaranteed by Theorem 2.9 in [7] and the convexity of the quadratic penalty. Examples of convex ρ include Huber's loss function and the cosh function. Motivated by the robustified likelihood in [14], the first term in (7) measures goodness-of-fit. The robustified likelihood is proportional to

$$p(\mathbf{y}|\mathbf{f}, \boldsymbol{\tau}) \propto \exp\left\{K\log|V| - \sum_{i=1}^{n}\rho(v_i'(\mathbf{y} - \mathbf{f}))\right\}, \tag{8}$$

where $K = E(z\psi(z))$, $\psi(z) = \rho^{(1)}(z)$ and z follows the standard normal distribution. The superscript $^{(1)}$ denotes the first derivative.

The ρ function depends on f only through the evaluation functional. By the same argument as in Section 5.1 of [7], the solution to (7) lies in a finite dimensional space and can be written as

$$f(\mathbf{t}) = \sum_{v=1}^{M} d_v\phi_v(\mathbf{t}) + \sum_{i=1}^{n} c_i\left(\sum_{k=1}^{q} R_k(\mathbf{t}_i, \mathbf{t})\right).$$

As a result, the minimization problem in (7) becomes

$$\min_{\mathbf{d},\mathbf{c}}\left\{\sum_{i=1}^{n}\rho(v_i'(\mathbf{y} - T\mathbf{d} - \Sigma_{\boldsymbol{\theta}}\mathbf{c})) + \frac{1}{2\sigma^2}\mathbf{c}'\Sigma_{\boldsymbol{\theta}}\mathbf{c}\right\}, \tag{9}$$

where $\mathbf{d} = (d_1, \ldots, d_M)'$ and $\mathbf{c} = (c_1, \ldots, c_n)'$.

There is no close-form solution to the above minimization problem due to the non-quadratic form of the ρ function. We propose to use the iteratively reweighted least squares algorithm with Fisher scoring for the optimization. We assume the first and second derivatives of the ρ function exist. Denote the current estimates of $\mathbf{d}$ and $\mathbf{c}$ as $\mathbf{d}_-$ and $\mathbf{c}_-$ respectively. Let $\mathbf{f}_- = T\mathbf{d}_- + \Sigma_{\boldsymbol{\theta}}\mathbf{c}_-$. The Fisher scoring algorithm iteratively solves the following penalized weighted least squares problem:

$$\min_{\mathbf{d},\mathbf{c}}\left\{(\mathbf{y}_- - T\mathbf{d} - \Sigma_{\boldsymbol{\theta}}\mathbf{c})'W_-(\mathbf{y}_- - T\mathbf{d} - \Sigma_{\boldsymbol{\theta}}\mathbf{c}) + \mathbf{c}'\Sigma_{\boldsymbol{\theta}}\mathbf{c}\right\}, \tag{10}$$

where $\mathbf{y}_- = \mathbf{f}_- - \sigma^2 W_-^{-1}\mathbf{u}_-$, $\mathbf{u}_- = \partial\sum_{i=1}^{n}\rho(v_i'(\mathbf{y} - \mathbf{f}))/\partial\mathbf{f}|_{\mathbf{f}=\mathbf{f}_-} = -\sum_{i=1}^{n}\psi(v_i'(\mathbf{y} - \mathbf{f}_-))v_i = -V\psi(V'(\mathbf{y} - \mathbf{f}_-))$, and $W_- = \sigma^2 E(\partial^2\sum_{i=1}^{n}\rho(v_i'(\mathbf{y} - \mathbf{f}))/\partial\mathbf{f}\partial\mathbf{f}'|_{\mathbf{f}=\mathbf{f}_-}) = \sigma^2\sum_{i=1}^{n}E(\psi^{(1)}(v_i'(\mathbf{y} - \mathbf{f}_-)))v_iv_i' = \sigma^2 V\mathrm{diag}(E(\psi^{(1)}(v_i'(\mathbf{y} - \mathbf{f}_-))))V'$. Here the expectation is taken under the current values of V and $\mathbf{f} = \mathbf{f}_-$. With V and $\mathbf{f}_-$ as the true values in the robustified

likelihood (8), $v_i'(\mathbf{y}-\mathbf{f}_-), i = 1,\ldots,n$ have a common distribution. The common distribution is the standard normal distribution when ρ is quadratic. Based on this observation, we further approximate $E(\psi^{(1)}(v_i'(\mathbf{y}-\mathbf{f}_-)))$ by $E(\psi^{(1)}(z))$. Therefore, $W_- = E(\psi^{(1)}(z))W$, i.e., we use the same correlation structure to model the working data and the observed data.

We call $\mathbf{y}_-$ the working data. Note that when the data are independent with a common variance, i.e., $W = I$, we have $\mathbf{y}_- = \mathbf{f}_- + \sigma\psi((\mathbf{y}-\mathbf{f}_-)/\sigma)/E(\psi^{(1)}(z))$. The working data differ from the empirical pseudo data in Oh et al. [18] only by the multiplying factor $E(\psi^{(1)}(z))$. Therefore, with a fixed smoothing parameter and scale parameter, the E–S algorithm can be thought of as a modified Fisher scoring algorithm with a step size $E(\psi^{(1)}(z))$. For example when $\rho(x) = x^2/2$ when $|x| \le c$ and $\rho(x) = c|x| - c/2$ otherwise (Huber's loss function), the step size is $P(|z| \le 1.345) = 0.82$ with $c = 1.345$.

4. Joint estimate of $\boldsymbol{\theta}$, $\boldsymbol{\tau}$, and σ^2

4.1. Indirect methods

Indirect methods select optimal smoothing parameters at each iteration of the iteratively reweighted penalized least squares algorithm. For generalized smoothing spline models, well-established smoothing parameter selection techniques such as GML, UBR and GCV can be used at each iteration because the working data are independent. For our robust model, the working data are correlated and involve unknown parameters $\boldsymbol{\tau}$; see (10). Choosing the optimal smoothing parameters based on the working data therefore needs techniques that allow for correlated data and preferably that also simultaneously estimate the covariance parameters. In the following, we chose to use the GML, GCV and U methods as shown in (4)–(6). Specifically, letting $\mathbf{y}_-$ be the current working data, the general procedure is as follows.

1. With the current values of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, obtain the estimate of $\mathbf{f}$ as $\hat{\mathbf{f}}_- = A_{W_-}\mathbf{y}_-$, where A_{W_-} is defined in (3) and $W_- = E(\psi^{(1)}(z))W$.
2. Update $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ by optimizing (4), (5), or (6). In this optimization, $\mathbf{y}$ is fixed at $\mathbf{y}_-$, W is replaced with $W_- = E(\psi^{(1)}(z))W$ which is a function of $\boldsymbol{\tau}$.
3. Obtain a robust estimate of σ^2. For instance, with $\mathbf{f}$ and $\boldsymbol{\tau}$ fixed at their current values, one can maximize the robustified likelihood (8) to obtain the estimate. Alternatively, one can resort to Huber's proposal II [13], i.e., minimize the following function with respect to σ

$$Q(\sigma, f) = \frac{1}{n}\sum_{i=1}^{n}\sigma\rho_H(v_i'(\mathbf{y}-\mathbf{f})) + \beta\sigma \tag{11}$$

   where ρ_H is the Huber's loss function, $\beta = E(z\psi(z) - \rho_H(z))$ and $\mathbf{f}$ is fixed at $\hat{\mathbf{f}}_-$. Note that v_i is the ith column of V which involves σ^2 since $VV' = \sigma^{-2}W$ where W is evaluated at the current $\boldsymbol{\tau}$.
4. Update the working data by $\hat{\mathbf{f}}_- - \sigma^2 W_-^{-1}\mathbf{u}_-$ with $\mathbf{u}_- = -V\psi(V'(\mathbf{y}-\hat{\mathbf{f}}_-))$.

The algorithm is iterated until estimates of the regression function, $\boldsymbol{\tau}$, and σ stabilize. Convergence is not guaranteed, however. As one can see from the algorithm, the GML, GCV, and U functions change from iteration to iteration. These changes in the objective functions along iterations make theoretical studies difficult. As noted in Gu [7], when it does converge, it converges to the fixed point of the Newton iteration with $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ optimally chosen by the GML, GCV, or U. Empirical studies suggest that the indirect methods are computationally efficient and converge in most situations. However, for generalized smoothing spline models, direct methods are proposed to overcome the convergence problem and improve effectiveness [29,31,16].

When data are independent with a common variance, i.e., $W = I$, the indirect algorithm is essentially the same as the E–S algorithm [18] with the added feature of joint scale estimation.

4.2. Direct methods: the rGML criterion

The GML method has been shown to be a reliable smoothing parameter selection criterion for both independent and correlated Gaussian data [5,26,8]. For general and generalized smoothing spline models, it is equivalent to the restricted maximum likelihood (REML) estimation method for variance components in mixed effects models. It is also equivalent to the marginal maximum likelihood estimation for hyper-parameters in an empirical Bayes model. In the following, we develop our rGML criterion from the latter point of view.

We assume the following prior for f,

$$F(\mathbf{t}) = \sum_{i=1}^{M} a_i\phi_i(\mathbf{t}) + \sigma Z(\mathbf{t}), \tag{12}$$

where $\mathbf{a} = (a_1,\ldots,a_M)' \overset{\text{i.i.d.}}{\sim} N(0,\xi)$, $Z(\mathbf{t})$ is a Gaussian process independent of $\mathbf{a}$ with $E(Z(\mathbf{t})) = 0$ and $E(Z(\mathbf{s})Z(\mathbf{t})) = \theta_1 R_1(\mathbf{s},\mathbf{t}) + \cdots + \theta_q R_k(\mathbf{s},\mathbf{t})$. Let $q(\mathbf{f}|\mathbf{a})$ be the Gaussian density of $\mathbf{f}$ with mean $T\mathbf{a}$ and covariance $\Sigma_{\boldsymbol{\theta}}$. The rGML estimates of

$\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ are defined as the maximizer of the marginal likelihood of $\mathbf{y}$ as $\xi \to \infty$,

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = \int p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) q(\mathbf{f}|\mathbf{a}) \mathrm{d}\mathbf{a} \mathrm{d}\mathbf{f}. \tag{13}$$

Note as $\xi \to \infty$, the prior on $\mathbf{a}$ goes to a noninformative prior. Let $\hat{\mathbf{f}}$ be the smoothing spline estimate at the convergence of the Fisher Scoring algorithm for fixed $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$. To approximate the above integral, we first use the second-order Taylor expansion of $l(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = -\log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2)$ around $\hat{\mathbf{f}}$. Let $\mathbf{y}_c$, $\mathbf{u}_c$, and W_c denote $\mathbf{y}_-$, $\mathbf{u}_-$, and W_- evaluated at $\hat{\mathbf{f}}$. The Taylor expansion leads to

$$l(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) \approx \frac{1}{2\sigma^2}(\mathbf{y}_c - \mathbf{f})' W_c (\mathbf{y}_c - \mathbf{f}) + C,$$

where $C = l(\mathbf{y}|\hat{\mathbf{f}}, \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) - \sigma^2 \mathbf{u}_c' W_c^{-1} \mathbf{u}_c/2$, which is independent of $\mathbf{f}$. In the above expansion, the Hessian matrix is replaced by W_c so it is robust and calculable. Also note that $\hat{\mathbf{f}}$, $\mathbf{y}_c$, $\mathbf{u}_c$, and W_c depend on $\boldsymbol{\theta}$, $\boldsymbol{\tau}$, and σ^2 implicitly. We suppress this dependence for simpler notation.

Let the QR decomposition of T be $T = (Q_1 Q_2)(R'\mathbf{0})'$. Following similar arguments to those in Liu et al. [15], the marginal likelihood of $\mathbf{y}$ (13), which we define as the rGML criterion, now reduces to

$$\mathrm{rGML}(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = C_1 \sigma^{-(n-M)} |Q_2'(\Sigma_{\boldsymbol{\theta}} + W_c^{-1}) Q_2|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{y}_c' Q_2 (Q_2'(\Sigma_{\boldsymbol{\theta}} + W_c^{-1}) Q_2)^{-1} Q_2' \mathbf{y}_c \right\}, \tag{14}$$

where $C_1 = \exp(-C)|W_c/\sigma^2|^{-1/2}$.

4.3. Direct methods: the rUBR criterion

As in [8], we measure the distance between the estimated regression function and the true regression function with the Kullback–Leibler distance. Let the likelihood of $\mathbf{y}$ be (8) with V_0 and f_0 as true values. We define the K–L distance as

$$L(f, f_0|\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = E\left\{ \frac{1}{n} \sum_{i=1}^{n} (\rho(v_i'(\mathbf{y} - \mathbf{f})) - \rho(v_{0i}'(\mathbf{y} - \mathbf{f}_0))) - \frac{K}{n} \log |V V_0^{-1}| \right\}. \tag{15}$$

Define $z_i = v_{0i}'(\mathbf{y} - \mathbf{f}_0)$ and $\mathbf{z} = (z_1, \ldots, z_n)' = V_0'(\mathbf{y} - \mathbf{f}_0)$. From the robustified likelihood (8), z_i's are independent and identically distributed. We again use the standard normal distribution to approximate the common distribution. We also assume that $\rho(z_i)$ and $\psi^{(1)}(z_i)$ are symmetrical functions and $\psi(z_i)$ is an odd function, which are satisfied by most robust loss functions. In Appendix A, we derived an approximately unbiased estimate of the risk function $R(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = E(L(\hat{f}, f_0|\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2))$ as follows

$$\mathrm{rUBR}(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = \frac{1}{n} l(\mathbf{y}|\hat{\mathbf{f}}) + \frac{E(\psi^2(z)) \mathrm{tr}\, A_{W_-}}{n E(\psi^{(1)}(z))} - \frac{E(\psi^{(1)}(z)\psi^2(z)) - E(\psi^{(1)}(z)) E(\psi^2(z))}{2n(E(\psi^{(1)}(z)))^2 / \mathrm{tr}\left(\mathrm{diag}^2(V^{-1} A_{W_-}' V)\right)}, \tag{16}$$

where $W_- = E(\psi^{(1)}(z)) W$.

When $\rho(x) = x^2/2$, it is can be checked that

$$\mathrm{rUBR}(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = \frac{1}{2n\sigma^2}(\mathbf{y} - \hat{\mathbf{f}})' W (\mathbf{y} - \hat{\mathbf{f}}) - \frac{1}{2n} \log |\sigma^{-2} W| + \frac{1}{n} \mathrm{tr}\, A_W.$$

Profiling out σ^2, Gu and Han [8] proved that the resulting function is asymptotically equivalent to (6) and consistent for estimating $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ (see Theorem 4.2 in their paper).

4.4. Direct methods: the rGCV for longitudinal data

The GCV method is well known for its optimal properties. In this section, we develop the direct robust GCV criterion for longitudinal data with $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^{-2}\mathrm{diag}(W_I)_{I=1}^N$ where N is the number of independent subjects and $\sigma^{-2} W_I$ is the covariance matrix for subject I. We develop the criterion based on the principle of cross validation. We did not do this for general correlated data because the definition of cross validation then becomes unclear.

Define the leave-one-out robust cross validation function as

$$\mathrm{rCV}(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = \frac{1}{n} \sum_{I=1}^{N} \sum_{j=1}^{n_I} \rho(v_{Ij}'(\mathbf{y}_I - \hat{f}^{-I}_{\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2}(\underline{\mathbf{t}}_I))) - \frac{K}{n} \sum_{I=1}^{N} \log(|V_I|), \tag{17}$$

where n_I is the number of observations for subject I, $\mathbf{y}_I$ is the response vector of the Ith subject, $\hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}$ is the smoothing spline estimate from (2) with fixed $\boldsymbol{\theta}$, $\boldsymbol{\tau}$, and σ^2 and the Ith subject deleted, $\underline{\mathbf{t}}_I$ is the covariate matrix for subject I, and v_{Ij} is the jth column of V_I where $V_I V_I' = \sigma^{-2} W_I$.

Straightforward calculation of (17) is time-consuming. In Appendix C, we derived a GCV approximation to (17) as follows:

$$\begin{aligned}\mathrm{rGCV}(\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2) &= \frac{1}{n} l(\mathbf{y}|\hat{\mathbf{f}}) + \frac{1}{nE(\psi^{(1)}(z))} \sum_{I=1}^{N} \psi(V_I'(\mathbf{y}_I - \hat{f}(\underline{\mathbf{t}}_I)))' V_I' (I - \bar{A}_{W_-,N})^{-1} \bar{A}_{W_-,N} (V_I')^{-1} \psi(V_I'(\mathbf{y}_I - \hat{f}(\underline{\mathbf{t}}_I))) \\ &\quad + \frac{1}{2n(E(\psi^{(1)}(z)))^2} \sum_{I=1}^{N} \psi(V_I'(\mathbf{y}_I - \hat{f}(\underline{\mathbf{t}}_I)))' V_I^{-1} \bar{A}'_{W_-,N} (I - \bar{A}'_{W_-,N})^{-1} \\ &\quad \times V_I \,\mathrm{Diag}(\psi^{(1)}(V_I'(\mathbf{y}_I - \hat{f}(\underline{\mathbf{t}}_I)))) V_I' (I - \bar{A}_{W_-,N})^{-1} \bar{A}_{W_-,N} (V_I')^{-1} \psi(V_I'(\mathbf{y}_I - \hat{f}(\underline{\mathbf{t}}_I))), \end{aligned} \tag{18}$$

where $\bar{A}_{W_-,N} = \sum_{I=1}^{N} A_{W_-,II}/N$, $A_{W_-,II}$ is the Ith block diagonal matrix of A_{W_-} and we assumed the matrix $I - \bar{A}_{W_-,N}$ is invertible. We suppressed the dependence of $\hat{\mathbf{f}}$ on $\boldsymbol{\theta}$, $\boldsymbol{\tau}$, and σ^2 for notational simplicity.

When data are independent with $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathrm{diag}(\sigma_i^{-2})$, the above robust GCV function reduces to

$$\begin{aligned}\mathrm{rGCV}(\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2) &= \frac{1}{n} l(\mathbf{y}|\hat{\mathbf{f}}) + \frac{\mathrm{tr}\, A_{W_-}}{nE(\psi^{(1)}(z))(n - \mathrm{tr}\, A_{W_-})} \sum_{i=1}^{n} \psi^2((y_i - \hat{f}_i)/\sigma_i) \\ &\quad + \frac{\mathrm{tr}^2 A_{W_-}}{2n(E(\psi^{(1)}(z)))^2 (n - \mathrm{tr}\, A_{W_-})^2} \sum_{i=1}^{n} \psi^2((y_i - \hat{f}_i)/\sigma_i) \psi^{(1)}((y_i - \hat{f}_i)/\sigma_i). \end{aligned} \tag{19}$$

5. Simulations

In this section, we evaluate the performances of the proposed methods in various settings in three simulation experiments. We first compare the direct rGML, rUBR, rGCV, indirect GML (denoted as iGML), indirect GCV (denoted as iGCV), and indirect U (denoted as iU) with independent errors. For this purpose, we generated data from model (1) with

$$f_0(t_i) = \sin(2\pi(1 - t_i)^2), \tag{20}$$

where t_i, $i = 1, \ldots, 100$ is a sequence of evenly spaced points in [0, 1]. In this simulation, four settings for independent ϵ_i are considered: $U(-1.8, 1.8)$, $N(0, 1)$, $0.9N(0, 1) + 0.1N(0.1, 10)$, and Cauchy (0, 0.6). We chose those to represent a range of short to long tailed distributions. We used the redescending Tukey's biweight $\psi = \rho'$ function throughout the simulations and applications, which is defined as $\psi(x) = x\left(1 - x^2/c^2\right)^2$ when $|x| < c$ and 0 otherwise. Here c is chosen to be 4.6851 which gives 95% asymptotic efficiency with respect to the standard normal distribution when the regression function is estimated parametrically.

Fig. 1 summarizes the simulation results based on 100 repetitions. Two metrics are used to compare among different methods, the mean squared error (MSE) and the KL distance defined respectively as

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (f(t_i) - f_0(t_i))^2,$$

$$\mathrm{KL} = \frac{1}{n\sigma^2} (\mathbf{f} - \mathbf{f}_0)' W (\mathbf{f} - \mathbf{f}_0) + \mathrm{tr}(\sigma_0^2 W W_0^{-1}/\sigma^2 - I) - \log|\sigma_0^2 W W_0^{-1}/\sigma^2|,$$

where f is the estimated regression function, and $\sigma_0^2 W_0^{-1}$ and $\sigma^2 W^{-1}$ are the true and estimated covariance matrices. The relative MSE and KL are the ratios of the MSE and KL of a method over those of the nonrobust GML method in (4). From Fig. 1, we observe that (1) the robust methods all perform better than the nonrobust GML when the error distribution has a longer tail than the standard normal, in terms of both MSE and KL. When the error distribution is uniform or normal, the nonrobust GML is more efficient. (2) The estimates of the regression function and scale parameter are accurate with all robust methods even for heavy tailed distributions, as reflected in the small MSE and KL values. (3) The direct and indirect approaches perform similarly in terms of MSE. The direct ones give slightly better KL. (4) The direct UBR method is good at estimating the regression function but not at the scale parameter.

The second simulation considers model (1) with correlated errors and

$$f_0(t_i) = 5 + 3\sin(2\pi t_i). \tag{21}$$

Two error models are considered: (1) $\epsilon_i = \gamma\epsilon_{i-1} + 0.3a_i$ and (2) $\epsilon_i = 0.3a_i + 0.3\gamma a_{i-1}$. In both error models, we consider $\gamma = -0.6$ and three distributions of a_i, $N(0, 1)$, $0.9N(0, 1) + 0.1N(0.1, 10)$, and Cauchy (0, 0.6) respectively.

Figs. 2 and 3 show the simulation results based on 100 repetitions. We observe that (1) robust methods outperform the nonrobust GML for longer tailed errors. With normal errors, the direct methods perform similarly to the nonrobust GML,

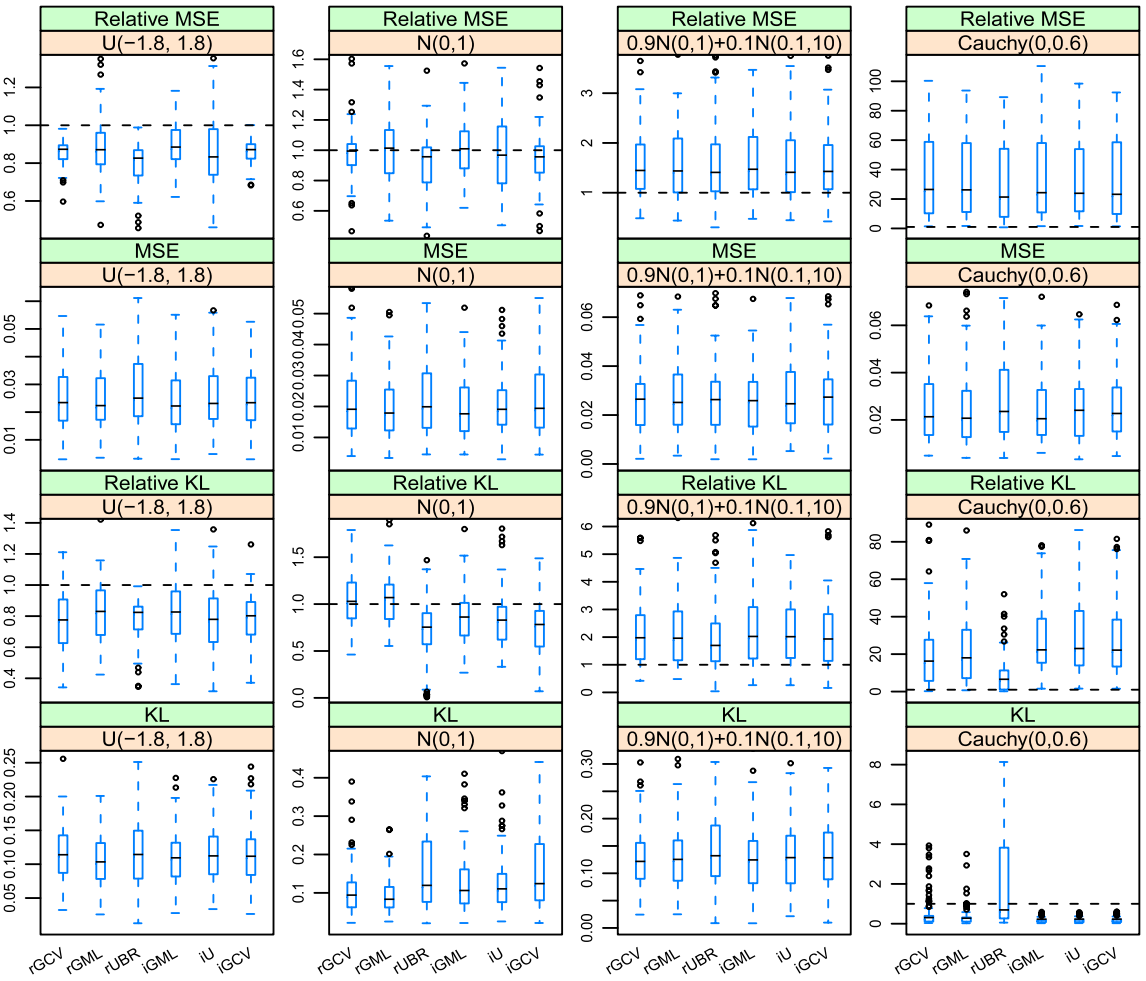

Fig. 1. Comparison among different methods for independent errors. A reference line at one has been added to the relative MSE and KL plots.

but the indirect ones are inferior in terms of the KL metric. (2) The direct methods generally outperform the indirect ones in terms of both MSE and KL. (3) Among indirect methods, the indirect U does notably worse than others for the MA(1) error model. This phenomenon was also reported in [8] with Gaussian errors, who noted an improved performance of the U method when the sample size increases. Fig. 3 shows the function estimates with the rGML method. The estimates are all fairly close to the truth.

In the third simulation, we consider model (1) for two groups of longitudinal data. Specifically, we consider the following model:

$$y_{gij} = f_0(g, t_{ij}) + b_{gi} + \sqrt{c_j}\epsilon_{gij}, \quad g = 1, 2,\ i = 1, \ldots, 10,\ j = 1, \ldots, 10, \tag{22}$$

where

$$f_0(g, t_{ij}) = 5I_{\{g=1\}} + (1 + 2I_{\{g=1\}}) \sin(2\pi t_{ij}), \tag{23}$$

and $t_{ij}, j = 1, \ldots, 10$ are evenly spaced points in $[0, 1]$. For each i, b_{gi}'s and ϵ_{gij}'s are mutually independent. Three different distributions of b_{gi} and ϵ_{gij} are considered: $N(0, 1)$, $0.9N(0, 1) + 0.1N(0.1, 10)$, and Cauchy $(0, 0.6)$. Therefore, a total of nine error distributions is considered in (22). To introduce heterogeneity, c_j is defined as equally spaced numbers from 0.5 to 3. In estimating model (23), we used a compound-symmetry structure for the covariance matrix, where the diagonal elements are a power series of $t_{ij}, j = 1, \ldots, 10$, with the power to be estimated.

Figs. 4 and 5 show the simulation results for the longitudinal data. Similar conclusions are reached as those from the correlated time series data. However, we note (1) the rUBR and indirect methods are less efficient when both the random effect and error are normally distributed, and (2) from the last row of Fig. 5 where there are outlying subjects, although the median estimates seem to be unbiased, the robust estimates showed sensitivity to subject-wise outliers.

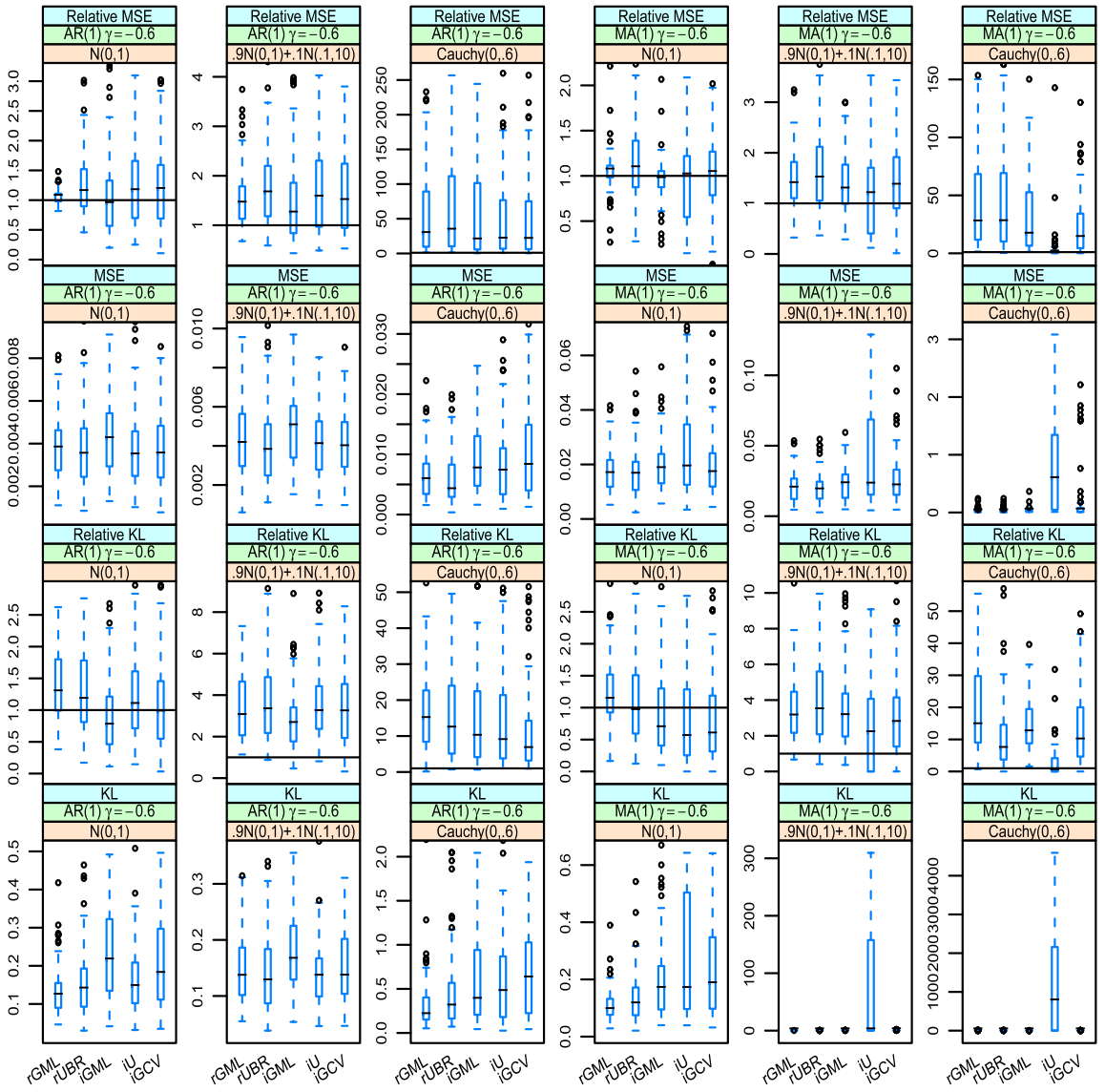

Fig. 2. Comparison among different methods with data generated from (21).

We have also performed simulations for other functions and sample sizes. In conclusion, the rGML and rGCV methods have the most stable performance among all settings and are superior in terms of the KL criterion. The direct UBR method is comparable and even slightly better in terms of the MSE for the regression function, however, its estimate of the covariance parameters is less reliable. Similar conclusions have been reached in [26] where GML, GCV, and UBR methods are developed for correlated Gaussian data. The indirect methods are computationally more efficient than the direct methods, but their performance is inferior when the errors are seriously correlated. While computational efficiency is important, it should be noted that a fit using the indirect methods took approximately 8 and 200 s for model (21) and (23) with contaminated normal errors, and one using the direct method took approximately 20 and 300 s. The timing is based on an unoptimized R program on a workstation with two AMD dualcore Opteron 2.0 GHz chips and 4 GB of RAM.

6. An application

We applied the proposed methods to the longitudinal CD4+ T-cell responses from twenty-four infected macaques in a Simian–Human Immunodeficiency Virus (SHIV89.6P) challenge study [4]. The goal of this analysis is to evaluate the protection against the depletion of CD4+ T-cells post SHIV89.6P infection, from the study vaccines. Each of the four priming-boosting vaccine regimens (1. DNA–DNA, 2. DNA–particle, 3. DNA-vaccinia virus, 4. vaccinia virus-DNA) was

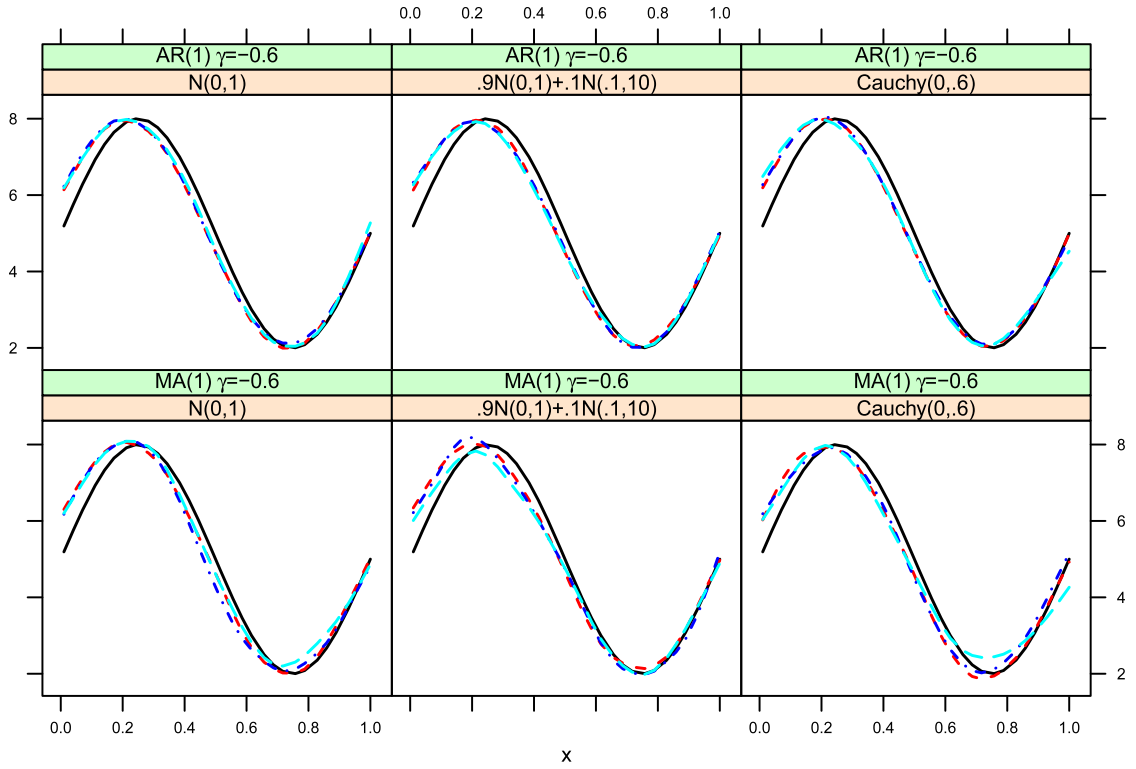

Fig. 3. Function estimates by the rGML method with data generated from (21). The solid line in each panel represents the true function, the short dashed line is the fit corresponding to the 5th percentile of the MSE, the long dashed line is the median fit, and the dot–dash line represents the fit corresponding to the 95th percentile of the MSE.

Table 1
Estimates of within and between subject standard deviations (sd) for group I: DNA/Vaccinia + Vaccinia/DNA and II: DNA/DNA + DNA/particle.

	Within subject sd		Between subject sd	
	Group I	Group II	Group I	Group II
rGML	183.83	226.59	413.04	200.60
GML	230.46	243.26	476.39	203.97

randomly assigned to six macaques. Three weeks after the last immunization, all animals were challenged with SHIV89.6P virus and infected afterwards. The observed CD4 trajectories show that most of the single-gene vaccinated (DNA–DNA or DNA–particle) macaques suffered from severe CD4 depletion in the first 3 weeks, while most animals receiving a combination of vaccinia virus and DNA vaccinations (DNA–Vaccinia or Vaccinia–Vaccinia) were protected from loss of CD4+ T cells over time.

Within both the single-gene-vaccinated and multigene-vaccinated groups, the CD4 responses in some animals showed distinct kinetics from the group average profile, either in terms of subject-specific trajectories or within-subject variability. Therefore we applied our robust methods: rGML, rUBR and rGCV. For comparison, we also fit the non-robust GML method to the data. Fig. 6 presents the observed CD4 counts (dotted lines) and the fitted CD4 curves from the day of challenge to the end of the study (all three robust fits were almost identical). Note that the rGML estimate appears to be very similar to the non-robust GML one for the unprotected group, which seems to show that the robust method is more sensitive to the subject-level outliers; however after the initial decrease in the first two weeks, the rGML estimate for the protected group backed up to a lower level than the non-robust GML estimate, which seems to fit the data better with the within-subject outliers presented. In the unprotected group, the CD4 curve had a rapid decline to around 200 and then stayed low; whereas after an initial drop in the CD4 counts the protected group had a rebound from around 600–750, then remained relatively in a steady state. A compound symmetric variance structure was used to model correlation within each group, which has two parameters: the within and between subject variances. Table 1 shows that the non-robust GML generally gives larger variance estimates, possibly due to effects of outliers. Our finding confirmed that the multigene regimens helped protect the macaques from disease.

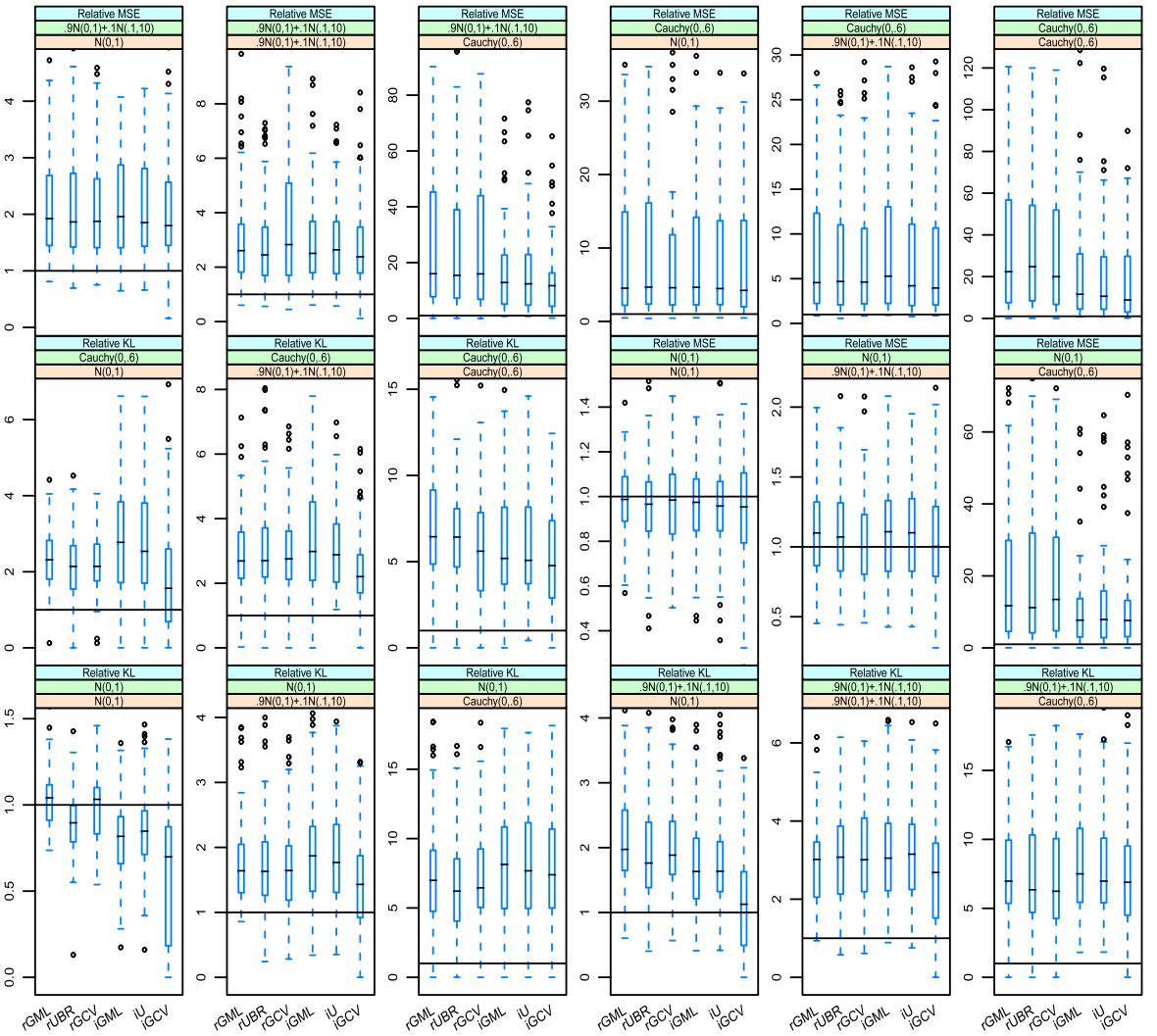

Fig. 4. Comparison among different methods for longitudinal data generated from (23). The middle strip is the distribution of the random effect and the bottom strip is the distribution of the error.

7. Discussion

In this paper, we defined a robust smoother for smoothing spline ANOVA models with correlated data. An iterative algorithm for solving the robust smoother was developed along with various robust smoothing parameter selection methods. We addressed the case when the correlation matrix can be parameterized, and we advocate joint estimates of the correlation parameters and the smoothing parameters. Indirect methods treat working data as normally distributed and utilize joint selection methods for correlated Gaussian data at each iteration. Direct methods including the rGML, rUBR, and rGCV are based on a robustified likelihood function. Simulations suggest that direct methods, although more computational costly than indirect methods, provide more accurate estimates. The computational requirements of the direct methods were not prohibitive in our examples.

Longitudinal immunological data from infectious diseases such as HIV/AIDS studies often have irregular kinetics that cannot be described well by parametric models. Due to the complexity of the infection process and the immune mechanism, some subject-specific trajectories may display large departures from the population trend, and some subjects may have outlying observations, both of which call for a robust nonparametric model. As has been shown by our real data example, the proposed methods work well in such situations.

Acknowledgments

Liu's research is partially supported by NSA H98230-09-1-0044. Qin's research is supported by a grant from NIAID. The authors thank the associate editor and a referee for their useful comments that have greatly improved this article.

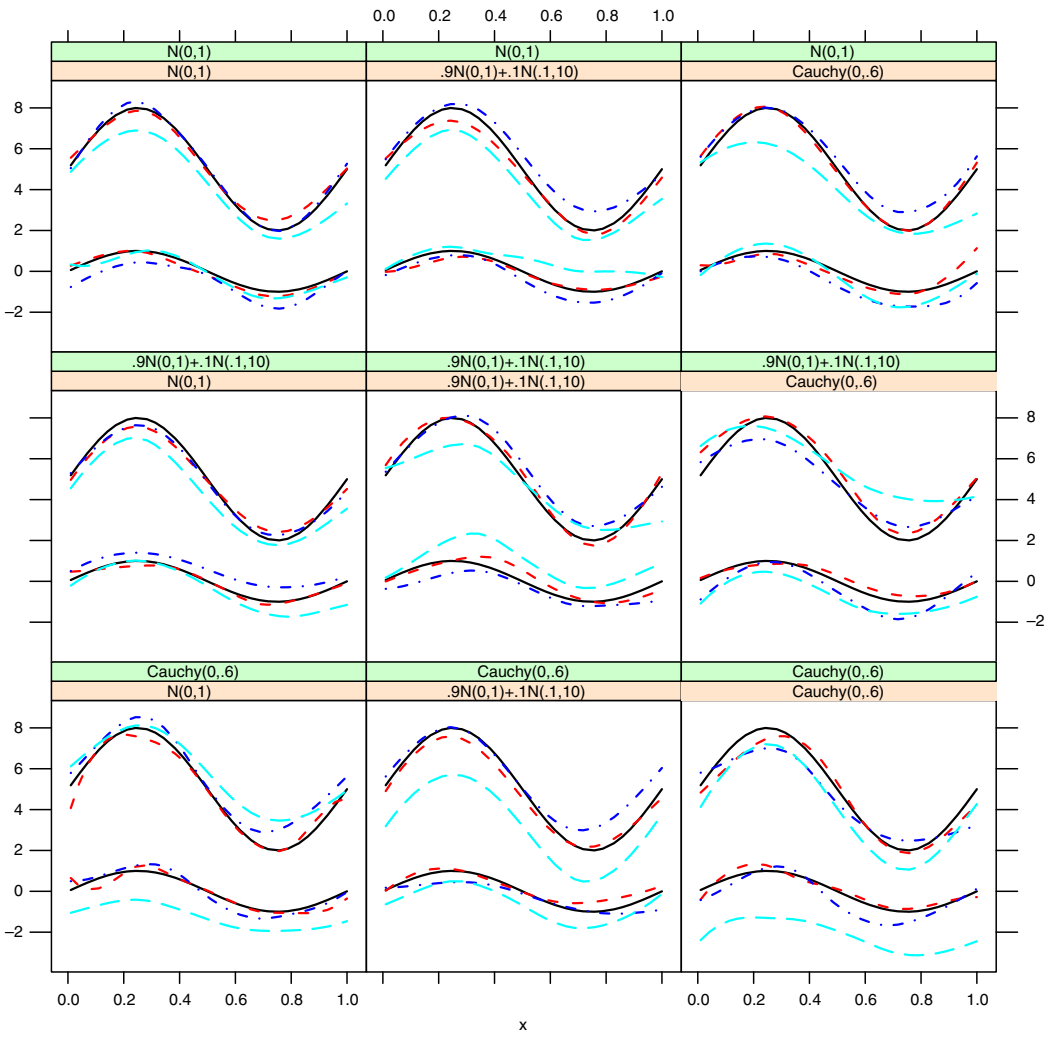

Fig. 5. Function estimates with rGML for longitudinal data generated from (22). The two groups of lines are the truth and fits for the functions $5+3\sin(2\pi x)$ and $\sin(2\pi x)$ respectively. The top strip is the distribution of the random effect and the bottom strip is the distribution of the error.

Appendix A. Derivation of the rUBR criterion

There is no close-form representation for $L(f, f_0|\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2)$. With the first order Taylor expansion, we have

$$\begin{aligned} L(f, f_0|\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) \approx E \Bigg\{ & -\frac{K}{n}\log|VV_0^{-1}| + \frac{1}{n}\sum_{i=1}^{n}\psi(\mathbf{z}_i)(v_i'(\mathbf{y}-\mathbf{f}) - v_{0i}'(\mathbf{y}-\mathbf{f}_0)) \\ & - \frac{1}{n}\psi(\mathbf{z})'V'(\mathbf{f}-\mathbf{f}_0) + \frac{1}{2n}\sum_{i=1}^{n}\psi^{(1)}(\mathbf{z}_i)(v_i'(\mathbf{y}-\mathbf{f}) - v_{0i}'(\mathbf{y}-\mathbf{f}_0))^2 \Bigg\} \\ = E \Bigg\{ & -\frac{K}{n}\log|VV_0^{-1}| + \frac{1}{n}\psi(\mathbf{z})'(V-V_0)'(\mathbf{y}-\mathbf{f}_0) \\ & + \frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\mathbf{y}-\mathbf{f}_0)'(V-V_0)\right) \\ & - \frac{1}{n}\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\mathbf{f}-\mathbf{f}_0)'V\right) + \frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'(\mathbf{f}-\mathbf{f}_0)(\mathbf{f}-\mathbf{f}_0)'V\right) \Bigg\}. \end{aligned}$$

It can be shown that $E(\psi(\mathbf{z}))'V'(\mathbf{f}-\mathbf{f}_0) = 0, E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\mathbf{f}-\mathbf{f}_0)'V\right)\right) = 0$, and $E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'(\mathbf{f}-\mathbf{f}_0)(\mathbf{f}-\mathbf{f}_0)'V\right)\right) = E(\psi^{(1)}(z))(\mathbf{f}-\mathbf{f}_0)'VV'(\mathbf{f}-\mathbf{f}_0)$. Now, the risk function $R(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = E(L(\hat{f}, f_0|\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2))$

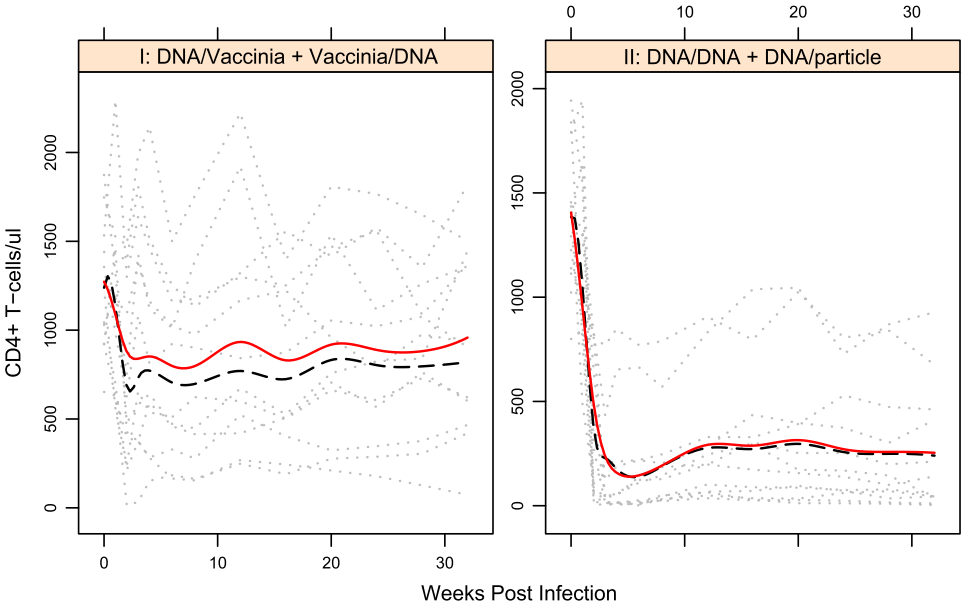

Fig. 6. Estimates of CD4 kinetics under two groups of treatments. The solid lines are with the nonrobust GML method and the dashed lines are with the rGML method.

is

$$R(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) \approx -\frac{K}{n}\log|VV_0^{-1}| + \frac{1}{n}E(\psi(\mathbf{z})'(V-V_0)'(\mathbf{y}-\mathbf{f}_0)) + \frac{1}{2n}E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\mathbf{y}-\mathbf{f}_0)'(V-V_0)\right)\right) + \frac{1}{2n}E(\psi^{(1)}(z))E\left((\hat{\mathbf{f}}-\mathbf{f}_0)'VV'(\hat{\mathbf{f}}-\mathbf{f}_0)\right). \tag{A.1}$$

Suppose that when we take the initial value of the Fisher scoring algorithm as $\mathbf{f}_0$, it converges in one iteration. Then we have $\hat{\mathbf{f}} \approx A_{W\_}(\mathbf{f}_0 - \sigma^2 W_{\_}^{-1}\mathbf{u})$, where $W_{\_} = \sigma^2 V\mathrm{diag}(E(\psi^{(1)}(v_i'(\mathbf{y}-\mathbf{f}_0))))V' \approx V\mathrm{diag}(E(\psi^{(1)}(z_i)))V' = E(\psi^{(1)}(z))W$, and $\mathbf{u} = -V\psi(V'(\mathbf{y}-\mathbf{f}_0)) \approx -V\psi(\mathbf{z})$. Here we assumed V is close to the truth V_0. As a result, we have $\hat{\mathbf{f}} \approx A_{W\_}\mathbf{f}_0 + A_{W\_}(V')^{-1}\psi(\mathbf{z})/E(\psi^{(1)})$. Now, for the last term in (A.1)

$$\begin{aligned} E\left((\hat{\mathbf{f}}-\mathbf{f}_0)'VV'(\hat{\mathbf{f}}-\mathbf{f}_0)\right) &= E\left(\left((A_{W\_}-I)\mathbf{f}_0 + A_{W\_}(V')^{-1}\psi(\mathbf{z})/E(\psi^{(1)}(z))\right)'VV'\left((A_{W\_}-I)\mathbf{f}_0 + A_{W\_}(V')^{-1}\psi(\mathbf{z})/E(\psi^{(1)}(z))\right)\right) \\ &= \mathbf{f}_0'(A_{W\_}-I)'VV'(A_{W\_}-I)\mathbf{f}_0 + E\left(\psi(\mathbf{z})'V^{-1}A_{W\_}'VV'(A_{W\_}-I)\mathbf{f}_0/E(\psi^{(1)}(z))\right) \\ &\quad + E\left(\mathbf{f}_0'(A_{W\_}-I)'VV'A_{W\_}(V')^{-1}\psi(\mathbf{z})/E(\psi^{(1)}(z))\right) \\ &\quad + E\left(\psi(\mathbf{z})'V^{-1}A_{W\_}'VV'A_{W\_}(V')^{-1}\psi(\mathbf{z})/(E(\psi^{(1)}(z)))^2\right) \\ &= \mathbf{f}_0'(A_{W\_}-I)'VV'(A_{W\_}-I)\mathbf{f}_0 + \frac{E(\psi^2(z))}{(E(\psi^{(1)}(z)))^2}\mathrm{tr}\left(V'A_{W\_}(V')^{-1}V^{-1}A_{W\_}'V\right). \end{aligned} \tag{A.2}$$

To approximate the first term above, consider a first-order Taylor expansion

$$\begin{aligned} \frac{1}{n}l(\mathbf{y}|\hat{\mathbf{f}}) &\approx -\frac{K}{n}\log|V| + \frac{1}{n}\sum_{i=1}^n \rho(z_i) + \frac{1}{n}\sum_{i=1}^n \psi(z_i)(v_i'(\mathbf{y}-\hat{\mathbf{f}}) - v_{0i}'(\mathbf{y}-\mathbf{f}_0)) \\ &\quad + \frac{1}{2n}\sum_{i=1}^n \psi^{(1)}(z_i)(v_i'(\mathbf{y}-\hat{\mathbf{f}}) - v_{0i}'(\mathbf{y}-\mathbf{f}_0))^2 \\ &= -\frac{K}{n}\log|V| + \frac{1}{n}\sum_{i=1}^n \rho(z_i) + \frac{1}{n}\psi(\mathbf{z})'\left((V-V_0)'(\mathbf{y}-\mathbf{f}_0)\right) \end{aligned}$$

$$-\frac{1}{n}\psi(\mathbf{z})'\left(V'(\hat{\mathbf{f}}-\mathbf{f}_0)\right)+\frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\mathbf{y}-\mathbf{f}_0)'(V-V_0)\right)$$
$$-\frac{1}{n}\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\hat{\mathbf{f}}-\mathbf{f}_0)'V\right)+\frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'(\hat{\mathbf{f}}-\mathbf{f}_0)(\hat{\mathbf{f}}-\mathbf{f}_0)'V\right). \quad (A.3)$$

Some algebra yields

$$E(\psi(\mathbf{z}))'\left(V'(\hat{\mathbf{f}}-\mathbf{f}_0)\right) \approx E(\psi(\mathbf{z}))'V'(A_{W_-}-I)\mathbf{f}_0+E(\psi(\mathbf{z}))'V'A_{W_-}(V')^{-1}\psi(\mathbf{z})/E(\psi^{(1)}(z))$$
$$= \frac{E(\psi^2(z))}{E(\psi^{(1)}(z))}\mathrm{tr}\,A_{W_-}, \quad (A.4)$$

$$E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'(\hat{\mathbf{f}}-\mathbf{f}_0)(\hat{\mathbf{f}}-\mathbf{f}_0)'V\right)\right)\approx E\left(\mathrm{tr}(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'(A_{W_-}-I)\mathbf{f}_0\mathbf{f}_0'(A_{W_-}-I)'V)\right)$$
$$+E\left(\mathrm{tr}(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'(A_{W_-}-I)\mathbf{f}_0\psi(\mathbf{z})'V^{-1}A_{W_-}'V)\right)/E(\psi^{(1)}(z))$$
$$+E\left(\mathrm{tr}(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'A_{W_-}(V')^{-1}\psi(\mathbf{z})\mathbf{f}_0'(A_{W_-}-I)'V)\right)/E(\psi^{(1)}(z))$$
$$+E\left(\mathrm{tr}(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'A_{W_-}(V')^{-1}\psi(\mathbf{z})\psi(\mathbf{z})'V^{-1}A_{W_-}'V)\right)/(E(\psi^{(1)}(z)))^2$$
$$=E(\psi^{(1)}(z))\mathbf{f}_0'(A_{W_-}-I)'VV'(A_{W_-}-I)\mathbf{f}_0$$
$$+E\left(\mathrm{tr}(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))V'A_{W_-}(V')^{-1}\psi(\mathbf{z})\psi(\mathbf{z})'V^{-1}A_{W_-}'V)\right)/(E(\psi^{(1)}(z)))^2$$
$$=E(\psi^{(1)}(z))\mathbf{f}_0'(A_{W_-}-I)'VV'(A_{W_-}-I)\mathbf{f}_0+\frac{E(\psi^2(z))}{E(\psi^{(1)}(z))}\mathrm{tr}(V'A_{W_-}(V')^{-1}V^{-1}A_{W_-}'V)$$
$$+\frac{E(\psi^{(1)}(z)\psi^2(z))-E(\psi^{(1)}(z))E(\psi^2(z))}{(E(\psi^{(1)}(z)))^2}\mathrm{tr}\left(\mathrm{diag}^2(V^{-1}A_{W_-}'V)\right), \quad (A.5)$$

and

$$E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)(\hat{\mathbf{f}}-\mathbf{f}_0)'V\right)\right)\approx E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)\mathbf{f}_0'(A_{W_-}-I)'V\right)\right)$$
$$+E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V-V_0)'(\mathbf{y}-\mathbf{f}_0)\psi(\mathbf{z})'V^{-1}A_{W_-}'V\right)\right)/E(\psi^{(1)}(z))$$
$$=E\left(\mathrm{tr}\left(\mathrm{diag}(\psi^{(1)}(\mathbf{z}))(V'(V_0')^{-1}-I)\mathbf{z}\psi(\mathbf{z})'V^{-1}A_{W_-}'V\right)\right)/E(\psi^{(1)}(z))$$
$$=E(z\psi(z))\mathrm{tr}\left((V'(V_0')^{-1}-I)V^{-1}A_{W_-}'V\right)$$
$$+\frac{E(z\psi^{(1)}(z)\psi(z))-E(\psi^{(1)}(z))E(z\psi(z))}{E(\psi^{(1)}(z))}\mathrm{tr}\left(\mathrm{diag}(V'(V_0')^{-1}-I)\mathrm{diag}(A_{W_-}'V)\right). \quad (A.6)$$

We ignore (A.6) based on the assumption that V and V_0 are close. Now combining (A.1)–(A.5), we have

$$\frac{1}{n}E(l(\mathbf{y}|\hat{\mathbf{f}}))\approx R(\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2)-\frac{K\log|V_0|}{n}+\frac{E(\rho(z))}{2}-\frac{E(\psi^2(z))\,\mathrm{tr}\,A_{W_-}}{nE(\psi^{(1)}(z))}$$
$$+\frac{E(\psi^{(1)}(z)\psi^2(z))-E(\psi^{(1)}(z))E(\psi^2(z))}{2n(E(\psi^{(1)}(z)))^2}\mathrm{tr}\left(\mathrm{diag}^2(V^{-1}A_{W_-}'V)\right).$$

Finally, a proxy of an unbiased risk estimator is proposed as

$$\mathrm{rUBR}(\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2)=\frac{1}{n}l(\mathbf{y}|\hat{\mathbf{f}})+\frac{E(\psi^2(z))\,\mathrm{tr}\,A_{W_-}}{nE(\psi^{(1)}(z))}-\frac{E(\psi^{(1)}(z)\psi^2(z))-E(\psi^{(1)}(z))E(\psi^2(z))}{2n(E(\psi^{(1)}(z)))^2/\mathrm{tr}\left(\mathrm{diag}^2(V^{-1}A_{W_-}'V)\right)}. \quad (A.7)$$

Appendix B. Leave-one-out lemma for longitudinal data

Lemma. *The minimizer* $f^{-L}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}$ *of*

$$\sum_{\substack{I=1\\I\neq L}}^{N}(\mathbf{y}_I-f(\underline{\mathbf{t}}_I))'W_I(\mathbf{y}_I-f(\underline{\mathbf{t}}_I))+\sum_{k=1}^{q}\theta_k^{-1}\|P_kf\|^2$$

is the minimizer of (2) *with* $\mathbf{y}=\tilde{\mathbf{y}}=(\mathbf{y}_1',\ldots,\mathbf{y}_{L-1}',f^{-L}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_L)',\mathbf{y}_{L+1}',\ldots,\mathbf{y}_N')'$.

The proof is very similar to that of Lemma 1 in [27] and is therefore omitted.

Appendix C. Derivation of the rGCV criterion

Similar to Gu and Xiang [9], we approximate the rCV function by

$$\mathrm{rCV}(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) = \frac{1}{n}\sum_{I=1}^{N}\sum_{j=1}^{n_I}\rho(v'_{Ij}(\mathbf{y}_I - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I))) - \frac{K}{n}\sum_{I=1}^{N}\log(|V_I|),$$

where $\hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}$ represents the smoothing spline estimate from (10) with the initial value being $\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}$ and the Ith subject deleted. We note that $\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}} = \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}$ because the Fisher Scoring algorithm converges to the right solution.

With the leave-one-out lemma applied to (10), we have $\hat{\mathbf{f}}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}} = A_{W\_}\tilde{\mathbf{y}}_-$, where $\tilde{\mathbf{y}}_-$ is the working data $\mathbf{y}_-$ with the Ith block component replaced by $\hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I)$. We also have $\hat{\mathbf{f}}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2} = \hat{\mathbf{f}}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}} = A_{W\_}\mathbf{y}_-$. Subtracting the two equations, we obtain

$$\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I) - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I) = A_{W\_,II}(\mathbf{y}_{-,I} - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I)),$$

where $A_{W\_,II}$ is the Ith block diagonal matrix of $A_{W\_}$. Similar to the GCV approximation to the CV function for independent data, we replace $A_{W\_,II}$ by $\bar{A}_{W\_,N} = \sum_{I=1}^{N} A_{W\_,II}/N$ and assume the matrix $I - \bar{A}_{W\_,N}$ is invertible. We thus obtain

$$\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I) - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I) = (I - \bar{A}_{W\_,N})^{-1}\bar{A}_{W\_,N}(\mathbf{y}_{-,I} - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)).$$

By the definition of working data, we have $\mathbf{y}_{-,I} - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I) = (V'_I)^{-1}\psi(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))/E(\psi^{(1)}(z))$. As a result,

$$\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I) - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I) = (I - \bar{A}_{W\_,N})^{-1}\bar{A}_{W\_,N}(V'_I)^{-1}\psi(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))/E(\psi^{(1)}(z)). \tag{C.1}$$

Now, with a second order Taylor expansion,

$$\begin{aligned}
\mathrm{rGCV}(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2) &= -\frac{K}{n}\sum_{I=1}^{N}\log(V_I) + \frac{1}{n}\sum_{I=1}^{N}\sum_{j=1}^{n_I}\rho(v'_{Ij}(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))\\
&\quad + \frac{1}{n}\sum_{I=1}^{N}\psi(v'_{Ij}(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))v'_{Ij}(\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I) - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I))\\
&\quad + \frac{1}{2n}\sum_{I=1}^{N}\sum_{j=1}^{n_I}\psi^{(1)}(v'_{Ij}(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))\left(v'_{Ij}(\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I) - \hat{f}^{-I}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}}(\underline{\mathbf{t}}_I))\right)^2\\
&= \frac{1}{n}l(\mathbf{y}|\hat{\mathbf{f}}) + \frac{1}{nE(\psi^{(1)}(z))}\sum_{I=1}^{N}\psi(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))'V'_I(I - \bar{A}_{W_,N})^{-1}\bar{A}_{W_,N}\\
&\quad \times (V'_I)^{-1}\psi(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I))) + \frac{1}{2n(E(\psi^{(1)}(z)))^2}\sum_{I=1}^{N}\psi(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I)))'\\
&\quad \times V_I^{-1}\bar{A}'_{W_,N}(I - \bar{A}'_{W_,N})^{-1}V_I\,\mathrm{Diag}(\psi^{(1)}(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I))))V'_I\\
&\quad \times (I - \bar{A}_{W_,N})^{-1}\bar{A}_{W_,N}(V'_I)^{-1}\psi(V'_I(\mathbf{y}_I - \hat{f}_{\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2}(\underline{\mathbf{t}}_I))).
\end{aligned}$$

References

[1] E. Cantoni, A robust approach to longitudinal data analysis, Canadian Journal of Statistics 32 (2004) 169–180.
[2] E. Cantoni, E. Ronchetti, Resistant selection of the smoothing parameter for smoothing splines, Statistics and Computing 11 (2001) 141–146.
[3] D. Cox, Asymptotics for M-type smoothing splines, Annals of Statistics 11 (1983) 530–551.
[4] N.A. Doria-Rose, C. Ohlen, P. Polacino, C.C. Pierce, M.T. Hensel, L. Kuller, T. Mulvania, D. Anderson, P.D. Greenberg, S.-L. Hu, N.L. Haigwood, Multigene dna priming-boosting vaccines protect macaques from acute CD4+-T-Cell depletion after simian-human immunodeficiency virus shiv89.6p mucosal challenge, Journal of Virology 77 (2003) 11563–11577.
[5] B. Efron, Selection criteria for scatterplot smoothers, Annals of Statistics 29 (2001) 470–505.
[6] P.J. Green, Iterative reweighted least squares for maximum likelihood estimates, and some robust and resistant alternatives, Journal of the Royal Statistical Society, Series B 46 (1984) 149–192.
[7] C. Gu, Smoothing Spline ANOVA Models, Springer, New York, 2002.
[8] C. Gu, C. Han, Optimal smoothing with correlated data, Sankhya: The Indian Journal of Statistics 70-A (2008) 38–72.
[9] C. Gu, D. Xiang, Cross-validating non-Gaussian data: generalized approximate cross-validation revisited, Journal of Computational and Graphical Statistics 10 (2001) 581–591.
[10] J.D. Hart, Kernel regression estimation with time series errors, Journal of the Royal Statistical Society, Series B 53 (1991) 173–187.

[11] X. He, W.-K. Fung, Z.-Y. Zhu, Robust estimation in generalized partial linear models for clustered data, Journal of the American Statistical Association 100 (2005) 1176–1184.
[12] P. Huber, in: E. Launer, G. Wilkinson (Eds.), Robust Smoothing, in: Robustness in Statistics, Academic Press, New York, 1979.
[13] P. Huber, Robust Statistics, John Wiley & Sons, 1981.
[14] R.M. Huggins, A robust approach to the analysis of repeated measures, Biometrics 49 (1993) 715–720.
[15] A. Liu, W. Meiring, Y. Wang, Testing generalized linear models using smoothing spline methods, Statistica Sinica 15 (2005) 235–256.
[16] A. Liu, T. Tong, Y. Wang, Smoothing spline estimation of variance functions, Journal of Computational and Graphical Statistics 16 (2007) 312–329.
[17] J. Mills, C. Field, D. Dupuis, Marginally specified generalized linear mixed models: a robust approach, Biometrics 58 (2002) 727–736.
[18] H.-S. Oh, D. Nychka, T. Lee, The role of pseudodata for robust smoothing with application to wavelet regression, Biometrika 94 (2007) 893–904.
[19] J. Opsomer, Y. Wang, Y. Yang, Nonparametric regression with correlated errors, Statistical Science 16 (2001) 134–153.
[20] J.S. Preisser, B.F. Qaqish, Robust regression for clustered data with application to binary responses, Biometrics 55 (1999) 574–579.
[21] G. Qin, Z. Zhu, Robustified maximum likelihood estimation in generalized partial linear mixed model for longitudinal data, Biometrics 65 (2009) 52–59.
[22] A.M. Richardson, Bounded influence estimation in the mixed linear model, Journal of the American Statistical Association 92 (1997) 154–161.
[23] A.M. Richardson, A.H. Welsh, Robust restricted maximum likelihood in mixed linear models, Biometrics 51 (1995) 1429–1439.
[24] S. Sinha, Robust analysis of generalized linear mixed models, Journal of the American Statistical Association 99 (2004) 451–460.
[25] J. Staudenmayer, E. Lake, M. Wand, Robustness for general design mixed models using the t-distribution, Statistical Modelling 9 (2009) 235–255.
[26] Y. Wang, Smoothing spline models with correlated random errors, Journal of the American Statistical Association 93 (1998) 341–348.
[27] Y. Wang, W. Guo, M.B. Brown, Spline smoothing for bivariate data with applications to association between hormones, Statistica Sinica 10 (2000) 377–397.
[28] Y.-G. Wang, X. Lin, M. Zhu, Robust estimating functions and bias correction for longitudinal data analysis, Biometrics 61 (2005) 684–691.
[29] D. Xiang, G. Wahba, A generalized approximate cross validation for smoothing splines with non-Gaussian data, Statistica Sinica 6 (1994) 675–692.
[30] K.K.W. Yau, A.Y.C. Kuk, Robust estimation in generalized linear mixed models, Journal of the Royal Statistical Society, Series B 64 (2002) 101–117.
[31] M. Yuan, Automatic smoothing for Poisson regression, Communications in Statistics—Theory and Methods 34 (2005) 603–617.