Generalized Nonparametric Mixed-Effect Models: Computation and Smoothing Parameter Selection

Chong Gu* *Purdue University, USA.*

Ping Ma *Harvard University, USA.*

SUMMARY

Generalized linear mixed-effect models are widely used for the analysis of correlated non-Gaussian data such as those found in longitudinal studies. In this article, we consider extensions with nonparametric fixed effects and parametric random effects. The estimation is through the penalized likelihood method, and our focus is on the efficient computation and the effective smoothing parameter selection. To assist efficient computation, the joint likelihood of the observations and the latent variables of the random effects is used instead of the marginal likelihood of the observations. For the selection of smoothing parameters and correlation parameters, direct cross-validation techniques are employed; the effectiveness of cross-validation with respect to a few loss functions are evaluated through simulation studies. Real data examples are also presented to illustrate some applications of the methodology.

KEYWORDS: Clustered data; Cross-validation; Longitudinal data; Mixed-effect model; Non-Gaussian regression; Penalized likelihood; Smoothing spline.

1 Introduction

Consider response data from exponential family distributions

$$Y_i \sim \exp\{(y\theta_i - b_i(\theta_i))/a(\phi) + c(y,\phi)\} \tag{1}$$

where θ is the canonical parameter, a (> 0), b, and c are known functions, and ϕ is the dispersion parameter. Given independent observations with covariates x, (x_i, Y_i), $i = 1, \ldots, n$, one may estimate θ as a function of x, $\theta_i = \theta(x_i)$; the dispersion is assumed to be a constant, either known or considered as a nuisance parameter. Assuming a linear model for some monotone transform ζ of θ, $\zeta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, one obtains a generalized linear model. See, e.g., McCullagh and Nelder (1989).

* *Address for correspondence*: Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.

When the responses Y_i are correlated, such as with longitudinal observations, one may model the correlation via random effects, and consider linear mixed-effect models of the form

$$\zeta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{b}, \tag{2}$$

where $\mathbf{x}_i^T \boldsymbol{\beta}$ are the fixed effects and $\mathbf{z}_i^T \mathbf{b}$ are the random effects, with the latent variable $\mathbf{b} \sim N(0, B)$. The unknown parameters are $\boldsymbol{\beta}$ and B, to be estimated from the data. See, e.g., Zeger and Karim (1991), Breslow and Clayton (1993), and McCulloch (1997).

In this article, we consider models of the form

$$\zeta_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b} \tag{3}$$

where the fixed effect $\eta(x)$ is assumed to be a smooth function on a generic domain $\mathcal{X}$, and the random effects $\mathbf{z}_i^T \mathbf{b}$ are as given in (2). Absent of the random effects $\mathbf{z}^T \mathbf{b}$, penalized likelihood regression has been studied extensively in the literature; see, e.g., Gu (2002, Chapter 5) for a comprehensive treatment of the subject. The purpose of this article is to study an approach to the estimation of $\eta(x)$ in (3) that allows one to take full advantage of the existing tools developed for independent data. Working with the joint likelihood of Y_i and $\mathbf{b}$, we minimize the penalized likelihood criterion

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\theta(\eta(x_i) + \mathbf{z}_i^T \mathbf{b}) - b(\theta(\eta(x_i) + \mathbf{z}_i^T \mathbf{b}))\} + \frac{1}{2n}\mathbf{b}^T \Sigma \mathbf{b} + \frac{\lambda}{2} J(\eta) \tag{4}$$

with respect to η and $\mathbf{b}$, where $\theta(\zeta)$ maps the modeling parameter to the canonical parameter, $\Sigma = a(\phi)B^{-1}$, $J(\eta)$ quantifies the roughness of η, and the smoothing parameter λ controls the trade-off between the goodness-of-fit and the smoothness of η. B is typically unknown yet we are not particularly concerned with its estimation; instead, we are more interested in the estimation of $\eta(x)$ or $\eta(x) + \mathbf{z}^T \boldsymbol{\beta}$ and will treat Σ as a tuning parameter like λ. An example of $J(\eta)$ is $\int_0^1 \ddot{\eta}^2 dx$ on $\mathcal{X} = [0, 1]$, which yields the popular cubic smoothing splines. To avoid the inconvenience of constrained optimization, the modeling parameter ζ should have the whole real line as its natural range.

A few examples are in order concerning the likelihood of Y and the random effects $\mathbf{z}^T \mathbf{b}$. Examples concerning $J(\eta)$ will be given in §2.

Example 1 (Logistic Regression) For $Y \sim \text{Bin}(m, p)$ with density $\binom{m}{y} p^y (1-p)^{m-y}$, $a(\phi) = 1$ is known, $\theta = \log\{p/(1-p)\}$, and $b(\theta) = m \log(1 + e^\theta)$. Setting $\zeta = \theta$, one has $\theta_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b}$. $\square$

Example 2 (Poisson Regression) For $Y \sim \text{Poisson}(\lambda)$ with density $\lambda^y e^{-\lambda}/y!$, $a(\phi) = 1$ is known, $\theta = \log\lambda$, and $b(\theta) = e^{\theta}$. Setting $\zeta = \theta$, one has $\theta_i = \eta(x_i) + \mathbf{z}_i^T\mathbf{b}$. The Poisson intensity λ is not to be confused with the smoothing parameter λ appearing in (4). $\Box$

Example 3 (Gamma Regression) For $Y \sim \text{Gamma}(\alpha, \beta)$ with density $\{\beta^{\alpha}\Gamma(\alpha)\}^{-1}y^{\alpha-1}e^{-y/\beta}$, $a(\phi) = \alpha^{-1}$, $\theta = -\mu^{-1}$, and $b(\theta) = \log(\mu)$, where $\mu = \alpha\beta = E[Y]$. Setting $\zeta = \log\mu$, one has $\theta_i = -\mu_i^{-1} = -\exp\{-(\eta(x_i) + \mathbf{z}_i^T\mathbf{b})\}$. $\Box$

These, along with the Gaussian regression treated in Gu and Ma (2003), shoud be broad enough to cover the lion's share of practical applications. Examples of correlation structures follow.

Example 4 (Longitudinal Observations) Consider a longitudinal study involving p subjects, where Y_i is taken from subject s_i with covariate x_i. Observations from different subjects are independent, while observations from the same subject are naturally correlated. The intra-subject correlation may be modeled by $\mathbf{z}_i^T\mathbf{b} = b_{s_i}$, where $\mathbf{b} \sim N(0, \sigma_s^2 I)$ and $\mathbf{z}_i$ is the s_i-th unit vector. The $p \times p$ matrix Σ involves only one tunable parameter. The random effects b_s can be interpreted as the subject effects, and we call them real. $\Box$

Example 5 (Clustered Observations) Consider observations from p clusters, such as in multicenter studies, 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 by $\mathbf{z}_i^T\mathbf{b} = b_{c_i}$, where $\mathbf{b} \sim N(0, B)$ with $B = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\mathbf{z}_i$ is the c_i-th unit vector. The $p \times p$ matrix Σ involves p tunable parameters on the diagonal. The random effects b_c are not quite interpretable in the setting, and we call them latent. $\Box$

As mentioned earlier, our primary concern is the estimation of $\eta(x) + \mathbf{z}^T\mathbf{b}$ or $\eta(x)$. With varying Σ and λ, the minimizer of (4) offers a family of candidate estimates, and an important practical issue is to achieve near optimal estimation precision through the proper selection of these tuning parameters. Cross-validation techniques developed for independent data can be used for the purpose, and simulation studies will be conducted to evaluate the empirical performance of cross-validation in the presence of random effects. Real data examples will also be presented to illustrate possible applications of the methodology.

The models of (3) were previously considered by Lin and Zhang (1999) and Karcher and Wang (2001), who tried to use the marginal likelihood of Y_i under the Bayes model of penalized likelihood estimation to select the tuning parameters, but had to settle with some quasi-likelihood approximations thereof. Lin and Zhang (1999) employed the Laplace approximation to integrate out $\mathbf{b}$, with the approximation accuracy largely beyond control, while Karcher and Wang (2001) chose to use

Markov chain Monte Carlo for the purpose, at a steep computational cost; quantitative evaluations of the empirical performances of the selection methods were also lacking.

The rest of the article is organized as follows. In §2, further details of the problem formulation are filled in, and in §3, the computational strategy is outlined. Cross-validation scores for the selection of the tuning parameters are discussed in §4, followed by the evaluation in §5 of their empirical performances through simulation studies. A couple of real data examples can be found in §6. Miscellaneous remarks are collected in §7.

2 Penalized Likelihood Regression

We shall now fill in some details concerning the roughness penalty $J(\eta)$ used in (3), and concerning penalized likelihood regression in general.

The minimization of (4) shall be performed in a space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a squre seminorm. The evaluation functional $[x]\eta = \eta(x)$ appears in the log likelihood term, and is assumed to be continuous in $\mathcal{H}$. A space $\mathcal{H}$ in which the evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot, \cdot)$, a non-negative definite function satisfying $R_x(\cdot) = R(x, \cdot) \in \mathcal{H}$, $\forall x \in \mathcal{X}$, and $\langle R(x, \cdot), f(\cdot)\rangle = f(x)$, $\forall f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. Typically, $\langle \cdot, \cdot \rangle = J(\cdot, \cdot) + \tilde{J}(\cdot, \cdot)$, where $J(\cdot, \cdot)$ is the semi inner product associated with $J(\cdot)$ and $\tilde{J}(\cdot, \cdot)$ is an inner product in the null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ of $J(\eta)$ when restricted therein. There exists a tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where the space $\mathcal{H}_J$ has $J(\eta)$ as its square norm and an RK R_J satisfying $J(R_J(x, \cdot), f(\cdot)) = f(x)$, $\forall f \in \mathcal{H}_J$. See, e.g., Gu (2002, §2.1).

Example 6 (Cubic Spline) For $x \in [0, 1]$, a choice of $J(\eta)$ is $\int_0^1 \ddot{\eta}^2 dx$, which yields the popular cubic splines. A choice of $\tilde{J}(f, g)$ is $(\int_0^1 f\,dx)(\int_0^1 g\,dx) + (\int_0^1 \dot{f}\,dx)(\int_0^1 \dot{g}\,dx)$, yielding $\mathcal{H}_J = \{\eta : \int_0^1 \eta dx = \int_0^1 \dot{\eta} dx = 0, J(\eta) < \infty\}$ and the RK $R_J(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(x_1 - x_2)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. The null space $\mathcal{N}_J$ has a basis $\{1, k_1(x)\}$, where $k_1(x) = x - 0.5$. See, e.g., Gu (2002, §2.3.3). □

Example 7 (Tensor Product Cubic Spline with Mixed Covariates) For $x = (x_{\langle 1\rangle}, x_{\langle 2\rangle}) \in [0, 1] \times \{1, \ldots, K\}$, one may decompose

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

where $\eta_\emptyset$ is a constant, $\eta_1(x_{\langle 1\rangle})$ is a function of $x_{\langle 1\rangle}$ satisfying $\int_0^1 \eta_1(x_{\langle 1\rangle})dx_{\langle 1\rangle} = 0$, $\eta_2(x_{\langle 2\rangle})$ is a function of $x_{\langle 2\rangle}$ satisfying $\sum_{x_{\langle 2\rangle}=1}^K \eta_2(x_{\langle 2\rangle}) = 0$, and $\eta_{1,2}(x_{\langle 1\rangle}, x_{\langle 2\rangle})$ satisfies $\int_0^1 \eta_{1,2}(x_{\langle 1\rangle}, x_{\langle 2\rangle})dx_{\langle 1\rangle} =$

0, $\forall x_{\langle 2\rangle}$, and $\sum_{x_{\langle 2\rangle}=1}^{K} \eta_{1,2}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = 0$, $\forall x_{\langle 1\rangle}$. One may use

$$J(\eta) = \theta_1^{-1} \int_0^1 (d^2\eta_1/dx_{\langle 1\rangle}^2)^2 dx_{\langle 1\rangle} + \theta_{1,2}^{-1} \int_0^1 \sum_{x_{\langle 2\rangle}=1}^{K} (d^2\eta_{1,2}/dx_{\langle 1\rangle}^2)^2 dx_{\langle 1\rangle},$$

where θ_1 and $\theta_{1,2}$ are extra smoothing parameters adjusting the relative weights of the roughness of different terms of $\eta(x)$. The null space $\mathcal{N}_J$ is of dimension $2K$ with a basis

$$\{1, k_1(x_{\langle 1\rangle}), I_{[x_{\langle 2\rangle}=j]} - 1/K, (I_{[x_{\langle 2\rangle}=j]} - 1/K)k_1(x_{\langle 1\rangle}), j = 1, \ldots, K-1\}.$$

The RK R_J in $\mathcal{H}_J$ is given by

$$\begin{aligned} R_J(x_1, x_2) &= \theta_1\{k_2(x_{1\langle 1\rangle})k_2(x_{2\langle 1\rangle}) - k_4(x_{1\langle 1\rangle} - x_{2\langle 1\rangle})\} \\ &\quad + \theta_{1,2}(I_{[x_{1\langle 2\rangle}=x_{2\langle 2\rangle}]} - 1/K)\{k_2(x_{1\langle 1\rangle})k_2(x_{2\langle 1\rangle}) - k_4(x_{1\langle 1\rangle} - x_{2\langle 1\rangle})\}; \end{aligned}$$

$f \in \mathcal{H}_J$ satisfies $\int_0^1 f(x_{\langle 1\rangle}, x_{\langle 2\rangle})dx_{\langle 1\rangle} = \int_0^1 \dot{f}(x_{\langle 1\rangle}, x_{\langle 2\rangle})dx_{\langle 1\rangle} = 0$, $\forall x_{\langle 2\rangle}$, and $\sum_{x_{\langle 2\rangle}=1}^{K} f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = 0$, $\forall x_{\langle 1\rangle}$. See, e.g., Gu (2002, §2.4.4). To force an additive model $\eta(x) = \eta_\emptyset + \eta_1(x_{\langle 1\rangle}) + \eta_2(x_{\langle 2\rangle})$, one may set $\theta_{1,2} = 0$ and remove $(I_{[x_{\langle 2\rangle}=j]} - 1/K)k_1(x_{\langle 1\rangle})$ from the null space basis. □

It is known that the minimizer of (4) in $\mathcal{H}$ resides in the space $\mathcal{N}_J \oplus \text{span}\{R_J(x_i, \cdot), i = 1, \ldots, n\}$. Gu and Kim (2002) considered the space $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(v_j, \cdot), j = 1, \ldots, q\}$, where $\{v_j\}$ is a random subset of $\{x_i\}$, and showed, in the absence of random effects, that the minimizer of (4) in $\mathcal{H}_q$ shares the same asymptotic convergence rates as that in $\mathcal{H}$, for $q \to \infty$ at rates much slower than n; for cubic splines, a rate of $q \asymp n^{2/9}$ may suffice. Without loss of generality, one may substitute an expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j \xi_j(x) \tag{5}$$

for $\eta(x)$ in (4), where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J$ and $\xi_j(x) = R_J(v_j, x)$; the "exact" minimizer in $\mathcal{H}$ is a special case with $q = n$. The computation is generally of the order $O(nq^2)$, so the restriction to $\mathcal{H}_q$ for some $q \ll n$ could greatly enhance the computational scalability of the technique.

The penalized likelihood estimation of (4) can be viewed as a Bayesian procedure, in which the roughness penalty $J(\eta)$ is proportional to the minus log likelihood of a prior on η; see, e.g., Wahba (1978, 1983) and Gu (1992b). For $\eta(x) \in \mathcal{H}_q$, the prior is given by $\eta(x) = \eta_0(x) + \eta_1(x)$ with η_0 diffuse in $\mathcal{H}_0$ and η_1 a Gaussian process with mean 0 and covariance function $E[\eta_1(x_1), \eta_1(x_2)] = \tau^2 R_J(x_1, \mathbf{v}^T)(R_J(\mathbf{v}, \mathbf{v}^T))^+ R_J(\mathbf{v}, x_2)$, independent of η_0, where $\mathbf{v} = (v_1, \ldots, v_q)^T$, $(\cdot)^+$ denotes the Moore-Penrose inverse, and $\tau^2 = a(\phi)/(n\lambda)$; see Kim and Gu (2003).

3 Computation

To carry out the penalized likelihood estimation of (4), we use two nested iterative loops. The inner loop minimizes (4) for fixed tuning parameters, and the outer loop selects the tuning parameters via the minimization of certain cross-validation scores. The inner loop calculations are discussed below; the outer loop cross-validation will be discussed in §4.

Fixing the smoothing parameter λ (and ones hidden in $J(\eta)$, if present) and the correlation parameters Σ, (4) may be minimized through Newton iteration. Write $l_i(\zeta_i) = -Y_i\theta(\zeta_i) + b(\theta(\zeta_i))$, $u_i = dl_i/d\zeta_i$, and $w_i = d^2 l_i/d\zeta_i^2$. The quadratic approximation of $l_i(\zeta_i)$ at the current estimate $\tilde{\zeta}_i$ is seen to be

$$l_i(\zeta_i) \approx l_i(\tilde{\zeta}_i) + \tilde{u}_i(\zeta_i - \tilde{\zeta}_i) + \tilde{w}_i(\zeta_i - \tilde{\zeta}_i)^2/2 = \tilde{w}_i(\tilde{Y}_i - \zeta_i)^2/2 + C_i,$$

where $\tilde{Y}_i = \tilde{\zeta}_i - \tilde{u}_i/\tilde{w}_i$ and C_i is independent of ζ_i; w_i is observed information, which is positive when $\zeta = \theta$, and may be replaced by the expected information when $\zeta \neq \theta$ to ensure positivity. The Newton iteration can thus be performed via iterated weighted least squares,

$$\sum_{i=1}^{n} \tilde{w}_i(\tilde{Y}_i - \eta(x_i) - \mathbf{z}_i^T\mathbf{b})^2 + \mathbf{b}^T\Sigma\mathbf{b} + n\lambda J(\eta). \tag{6}$$

Substituting (5) into (6), the numerical problem becomes the minimization of

$$(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c} - Z\mathbf{b})^T\tilde{W}(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c} - Z\mathbf{b}) + \mathbf{b}^T\Sigma\mathbf{b} + n\lambda\mathbf{c}^TQ\mathbf{c} \tag{7}$$

with respect to $\mathbf{d}$, $\mathbf{c}$, and $\mathbf{b}$, where $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$, S is $n \times m$ with the (i,ν)th entry $\phi_\nu(x_i)$, R is $n \times q$ with the (i,j)th entry $\xi_j(x_i)$, Z is $n \times p$ with the ith row $\mathbf{z}_i^T$, Q is $q \times q$ with the (j,k)th entry $J(\xi_j, \xi_k) = R_J(v_j, v_k)$, and $\tilde{W} = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$. The solution of (7) satisfies the normal equation

$$\begin{pmatrix} S_w^TS_w & S_w^TR_w & S_w^TZ_w \\ R_w^TR_w & R_w^TR_w + (n\lambda)Q & R_w^TZ_w \\ Z_w^TS_w & Z_w^TR_w & Z_w^TZ_w + \Sigma \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} S_w^T\tilde{\mathbf{Y}}_w \\ R_w^T\tilde{\mathbf{Y}}_w \\ Z_w^T\tilde{\mathbf{Y}}_w \end{pmatrix}, \tag{8}$$

where $S_w = \tilde{W}^{1/2}S$, $R_w = \tilde{W}^{1/2}R$, $Z_w = \tilde{W}^{1/2}Z$, and $\tilde{\mathbf{Y}}_w = \tilde{W}^{1/2}\tilde{\mathbf{Y}}$. The normal equation of (8) can be solved by a Cholesky decomposition followed by backward and forward substitutions. Possible singularity of the matrix can be properly handled through pivoting in Cholesky decomposition; see, e.g., Kim and Gu (2003) for details.

On the convergence of Newton iteration, the "fitted values" $\hat{\mathbf{Y}}_w = S_w\mathbf{d} + R_w\mathbf{c} + Z_w\mathbf{b}$ of (6) can be written as $\hat{\mathbf{Y}}_w = A_w(\lambda, \Sigma)\tilde{\mathbf{Y}}_w$, where the diagonals of the smoothing matrix

$$A_w(\lambda, \Sigma) = (S_w, R_w, Z_w) \begin{pmatrix} S_w^T S_w & S_w^T R_w & S_w^T Z_w \\ R_w^T R_w & R_w^T R_w + (n\lambda)Q & R_w^T Z_w \\ Z_w^T S_w & Z_w^T R_w & Z_w^T Z_w + \Sigma \end{pmatrix}^+ \begin{pmatrix} S_w^T \\ R_w^T \\ Z_w^T \end{pmatrix}$$

will be needed in the cross-validation of §4 for the selection of tuning parameters.

Under the Bayes model of penalized likelihood regression, (6) is of the form of a Gaussian joint log likelihood of $\tilde{Y}_i$, $\mathbf{b}$, and η. On the convergence of Newton iteration, $\hat{\eta}(x) + \mathbf{z}^T\hat{\mathbf{b}}$ can be shown to be the posterior mean of $\eta(x) + \mathbf{z}^T\mathbf{b}$ under the setting, which, combined with the corresponding posterior standard deviation, yields the Bayesian confidence interval of Wahba (1983). Detailed formulas and derivations follow straightforward modifications of those with independent data, to be found in, e.g., Wahba (1983), Gu (1992b), and Kim and Gu (2003).

4 Cross-Validation

For the selection of the tuning parameters in (4), one may calculate cross-validation scores for estimates with fixed tuning parameters, and employ standard optimization algorithms such as those in Dennis and Schnabel (1996) to minimize the cross-validation scores as functions of the tuning parameters. We shall discuss below the cross-validation scores to use in logistic regression, Poisson regression, and gamma regression, and evaluate in §5 their empirical performances through simulation studies.

Using a general method derived in Gu and Xiang (2001), a generalized approximate cross-validation score is given by

$$V_g(\lambda, \Sigma) = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i\hat{\theta}_i - b_i(\hat{\theta}_i)\} + \frac{\mathrm{tr}(A_w\tilde{W}^{-1})}{n - \mathrm{tr}A_w}\frac{1}{n}\sum_{i=1}^{n}Y_i(d\theta_i/d\zeta_i|_{\hat{\zeta}_i})(-\hat{u}_i), \tag{9}$$

where θ_i, $d\theta_i/d\zeta_i$, and u_i are evaluated at the minimizer of (4) with fixed tuning parameters, and A_w and $\tilde{W}$ are as given in §3 evaluated on the convergence of Newton iteration; see Kim (2003). The cross-validation score of (9) targets the Kullback-Leibler loss

$$\mathrm{KL}(\theta, \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\{\mu_i\,(\theta_i - \hat{\theta}_i) - (b(\theta_i) - b(\hat{\theta}_i))\}, \tag{10}$$

where $\mu_i = db/d\theta_i = E[Y_i]$. The empirical performance of (9) is outstanding for logistic regression with Bernoulli data in the simulations of Xiang and Wahba (1996) and Gu and Xiang (2001), in the absence of random effects.

For the logistic regression of Example 1, $d\theta/d\zeta = 1$, $u_i = m_i p_i - Y_i$, and $w_i = m_i p_i(1-p_i)$. When binomial data $Y_i \sim \text{Bin}(m_i, p_i)$ with $m_i > 1$ are perceived as grouped Bernoulli data, the verbatim application of (9) amounts to "delete-m" instead of "delete-1" cross-validation, however. The "delete-1" version can be shown to be

$$V(\lambda, \Sigma) = -\frac{1}{N}\sum_{i=1}^{n}\{Y_i \log \hat{p}_i + (m_i - Y_i)\log(1-\hat{p}_i)\} + \frac{\text{tr}(A_w M \tilde{W}^{-1})}{N - \text{tr}A_w}\frac{1}{N}\sum_{i=1}^{n} Y_i(1-\hat{p}_i), \qquad (11)$$

where $N = \sum_{i=1}^{n} m_i$ and $M = \text{diag}(m_1, \dots, m_n)$; see Kim (2003).

For the Poisson regression of Example 2, $d\theta/d\zeta = 1$, $u_i = e^{\theta_i} - Y_i$, and $w_i = e^{\theta_i}$. In the simulations of Kim (2003) absent of the random effects, the empirical performance of (9) for Poisson regression is generally acceptable, but not as good as that of a "delete-1-count" cross-validation score derived through a density estimation interpretation of Poisson regression. The "delete-1-count" cross-validation score is given by

$$V(\lambda, \Sigma) = -\frac{1}{N}\sum_{i=1}^{n}\{Y_i\hat{\theta}_i - e^{\hat{\theta}_i}\} + \alpha\frac{\text{tr}(P_y \tilde{R} H^{+} \tilde{R}^T P_y^T)}{N(N-1)}, \qquad (12)$$

where $\alpha \geq 1$ is a constant, $N = \sum_{i=1}^{n} Y_i$, $\tilde{R} = (S, R, Z)$, $P_y = (I - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^T/N)\text{diag}(\tilde{\mathbf{y}})$ with $\tilde{\mathbf{y}} = (\sqrt{Y_1}, \dots, \sqrt{Y_n})^T$, and

$$H = \begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} & V_{\phi,z} \\ V_{\xi,\phi} & V_{\xi,\xi} + (n\lambda)Q/N & V_{\xi,z} \\ V_{z,\phi} & V_{z,\xi} & V_{z,z} + \Sigma/N \end{pmatrix}$$

with $V_{\xi,\phi}$ a $q \times m$ matrix having the (j,ν)th entry

$$\frac{1}{N}\sum_{i=1}^{n} e^{\hat{\theta}_i}\xi_j(x_i)\phi_\nu(x_i) - \frac{1}{N}\sum_{i=1}^{n} e^{\hat{\theta}_i}\xi_j(x_i)\frac{1}{N}\sum_{i=1}^{n} e^{\hat{\theta}_i}\phi_\nu(x_i)$$

and other V matrices similarly defined; it is known that $\sum_{i=1}^{n} e^{\hat{\theta}_i} = N$. See Gu and Wang (2003) and Kim (2003).

For the gamma regression of Example 3, $d\theta_i/d\zeta_i = 1/\mu_i$, $u_i = 1 - Y_i/\mu_i$, and $w_i = Y_i/\mu_i$. A slight modification of (9) yields

$$V(\lambda, \Sigma) = -\frac{1}{n}\sum_{i=1}^{n}\{-Y_i/\hat{\mu}_i - \log\hat{\mu}_i\} + \alpha\frac{\text{tr}(A_w \tilde{W}^{-1})}{n - \text{tr}A_w}\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_i/\hat{\mu}_i - 1)/\hat{\mu}_i, \qquad (13)$$

where $\alpha \geq 1$ is a constant.

The cross-validation scores of (11), (12), and (13) are very effective in the absence of random effects, as shown in the empirical studies of Kim (2003); an α in the range of $1.2 \sim 1.4$ in (12) and (13) helps to prevent occasional severe undersmoothing typically suffered by cross-validation methods, with little loss of general effectiveness.

5 Empirical Performance of Cross-Validation

We now evaluate the empirical performances of the cross-validation scores (11), (12), and (13) in their respective settings through simulation studies.

5.1 Loss Functions

To assess the performance of $\hat{\theta}_i$ as estimates of θ_i, one may use the symmetrized Kullback-Leibler loss

$$L_1(\lambda, \Sigma) = \mathrm{KL}(\theta, \hat{\theta}) + \mathrm{KL}(\hat{\theta}, \theta) = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{\mu}_i)(\theta_i - \hat{\theta}_i), \tag{14}$$

and the associated weighted mean square error in ζ,

$$\tilde{L}_1(\lambda, \Sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{d\mu_i}{d\zeta_i}\frac{d\theta_i}{d\zeta_i}(\zeta_i - \hat{\zeta}_i)^2, \tag{15}$$

where the dependence of the losses on the tuning parameters (through the estimates) are made explicit in the notation. For the logistic regression of Example 1, $(d\mu_i/d\zeta_i)(d\theta_i/d\zeta_i) = mp_i(1-p_i) = w_i$, for the Poisson regression of Example 2, $(d\mu_i/d\zeta_i)(d\theta_i/d\zeta_i) = e^{\theta_i} = w_i$, and for the gamma regression of Example 3, $(d\mu_i/d\zeta_i)(d\theta_i/d\zeta_i) = 1 = E[w_i]$. When the random effects $\mathbf{z}_i^T\mathbf{b}$ are real, $L_1(\lambda, \Sigma)$ and $\tilde{L}_1(\lambda, \Sigma)$ are among natural losses for the estimation of $\zeta_i = \eta(x_i) + \mathbf{z}_i^T\mathbf{b}$ by $\hat{\zeta}_i = \hat{\eta}(x_i) + \mathbf{z}_i^T\hat{\mathbf{b}}$, and (15) can be written as

$$\tilde{L}_1(\lambda, \Sigma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} + Z(\hat{\mathbf{b}} - \mathbf{b}))^T W(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} + Z(\hat{\mathbf{b}} - \mathbf{b})), \tag{16}$$

where $\boldsymbol{\eta} = (\eta(x_1), \ldots, \eta(x_n))^T$ and $W = \mathrm{diag}((d\mu_1/d\zeta_1)(d\theta_1/d\zeta_1), \ldots, (d\mu_n/d\zeta_n)(d\theta_n/d\zeta_n))$.

When the random effects $\mathbf{z}_i^T\mathbf{b}$ are latent, the loss functions involving $Z\mathbf{b}$ may not make much practical sense. Replacing $y\theta(\zeta) - b(\theta(\zeta))$ in (1) by its quadratic approximation at some $\tilde{\zeta}$, the approximate likelihood is Gaussian with pseudo observation $\tilde{y} = \tilde{\zeta} - \tilde{u}/\tilde{w}$, mean $\zeta = \eta(x) + \mathbf{z}^T\mathbf{b}$, and variance $a(\phi)/\tilde{w}$, where u and w are as defined in §3, and in turn the "marginal likelihood" of $\sqrt{\tilde{w}_i}\,\tilde{Y}_i$ is Gaussian with mean $\sqrt{\tilde{w}_i}\,\eta(x_i)$ and covariance $a(\phi)I + Z_wBZ_w^T$. Similar to the parallel derivation in Gu and Ma (2003, §4), one has

$$
\begin{aligned}
L_2(\lambda,\Sigma) &= \frac{1}{n}(\hat{\boldsymbol{\eta}}_w-\boldsymbol{\eta}_w)^T P^{\perp}_{Z_w}(\hat{\boldsymbol{\eta}}_w-\boldsymbol{\eta}_w) \\
&\propto \frac{1}{n}(\hat{\boldsymbol{\eta}}_w-\boldsymbol{\eta}_w)^T P^{\perp}_{Z_w}(a(\phi)I+Z_wBZ_w^T)^{-1}P^{\perp}_{Z_w}(\hat{\boldsymbol{\eta}}_w-\boldsymbol{\eta}_w),
\end{aligned}
\tag{17}
$$

where $\boldsymbol{\eta}_w=\tilde{W}^{1/2}\boldsymbol{\eta}$, Z_w is as in (8), and $P^{\perp}_{Z_w}=I-Z_w(Z_w^TZ_w)^+Z_w^T$, which makes an adequate loss function for the assessment of the estimation of $P^{\perp}_{Z_w}\boldsymbol{\eta}_w$ by $P^{\perp}_{Z_w}\hat{\boldsymbol{\eta}}_w$; the projection ensures the identifiability of the target function and the loss is independent of B. For definiteness, one may substitute the $\tilde{W}$ in (17) by the W in (16).

With a mixture of real and latent random effects, one may 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. A loss function for the estimation can then be defined by

$$
L_3(\lambda,\Sigma)=\frac{1}{n}(\hat{\boldsymbol{\eta}}_w-\boldsymbol{\eta}_w+Z_{w1}(\hat{\mathbf{b}}_1-\mathbf{b}_1))^T P^{\perp}_{Z_{w2}}(\hat{\boldsymbol{\eta}}_w-\boldsymbol{\eta}_w+Z_{w1}(\hat{\mathbf{b}}_1-\mathbf{b}_1)),
\tag{18}
$$

where $Z_w=(Z_{w1},Z_{w2})$.

5.2 Simulations with Real Random Effects

Consider $\zeta_i=\eta(x_i)+b_{s_i}$, $i=1,\ldots,200$, where

$$
\eta(x)=1980\,x^7(1-x)^3+858\,x^2(1-x)^{10}-2
$$

is a bimodal beta mixture, $\{x_i\}$ is a random sample from $U(0,1)$, $s_i\in\{1,\ldots,20\}$, 10 each, and $b_s\sim N(0,0.5^2)$; the random effects are as in Example 4. Bernoulli (binomial with $m_i=1$) data with logit ζ_i, Poisson data with log intensity ζ_i, and gamma data with log mean ζ_i and dispersion $a(\phi)=1$ were generated. The cubic splines of Example 6 were used in the penalized likelihood estimation through (4).

For each of the three distribution families, one hundred replicates of samples were generated, and estimates were calculated with the tuning parameters (λ_v,Σ_v) minimizing (11), (12), or (13); for (12) and (13), calculations were done with $\alpha=1,1.2,1.4,1.6,1.8$. Estimates were also calculated with the tuning parameters (λ_m,Σ_m) minimizing the losses $L_1(\lambda,\Sigma)$ of (14) and $\tilde{L}_1(\lambda,\Sigma)$ of (16).

The simulation results are summarized in Figures 1 and 2. In the frames of Figures 1, the losses $L_1(\lambda_v,\Sigma_v)$ of some of the cross-validated estimates are plotted against the best achievable losses $L_1(\lambda_m,\Sigma_m)$ given the data; a point on the dotted diagonal lines indicates a perfect performance by cross-validation. Plots with $\tilde{L}_1$ are similar. In the left frame of Figures 2, the relative efficacy

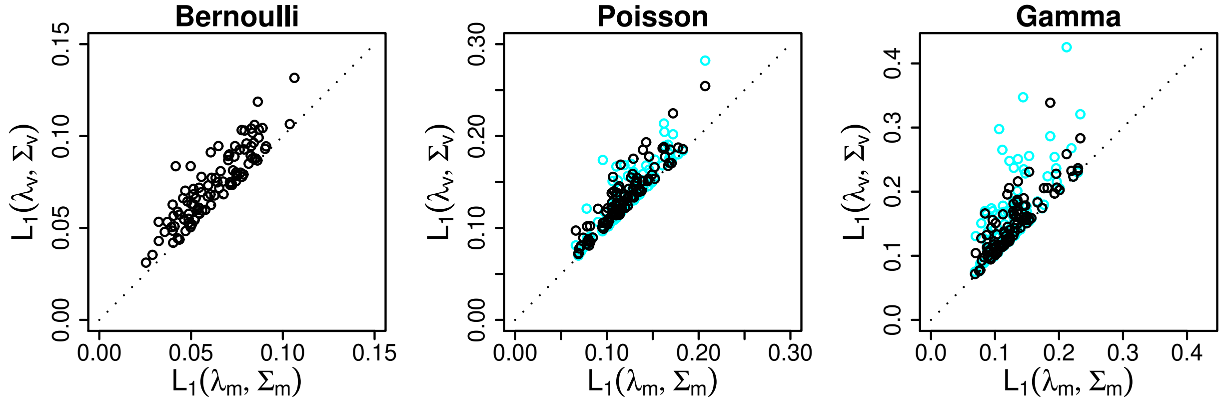

Figure 1: Simulations with Real Random Effects. Left: Performance of (11) for Bernoulli data. Center: Performance of (12) for Poisson data. Right: Performance of (13) for gamma data. For Poisson and gamma data, results of $\alpha = 1$ are in faded circles and those of $\alpha = 1.4$ are in circles.

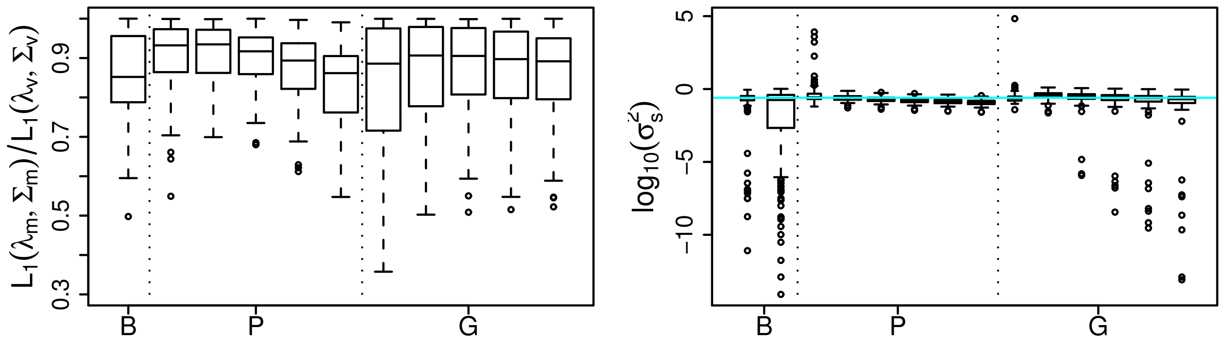

Figure 2: Simulation with Real Random Effects. Left: Relative efficacy for Bernoulli (B), Poisson (P), and gamma (G) data; for P and G, the plots are for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, in order. Right: Variance σ_s^2 "estimated" through Σ_m (thinner boxes) and Σ_v (fatter boxes) for Bernoulli (B), Poisson (P), and gamma (G) data; faded line is true $\sigma_s^2 = 0.25$.

$L_1(\lambda_m, \Sigma_m)/L_1(\lambda_v, \Sigma_v)$ are summarized in box-plots. The right frame plots the variance σ_s^2 "estimated" through the variance ratio $a(\phi)/\sigma_s^2$ in Σ_m and Σ_v, which is highly unreliable; note that the cross-validation scores are designed to minimize the loss for the estimation of $\eta(x) + \mathbf{z}^T\mathbf{b}$ but not σ_s^2. It is seen that in the Bernoulli simulation, many Σ_m effectively suppress the random effects b_s, while in the Poisson and gamma simulations, quite a few Σ_m leave b_s unpenalized.

5.3 Simulations with Latent Random Effects

For latent random effects, we keep the settings of §5.2 but replace b_{s_i} by $b_{c_i} \in \{1, 2\}$, 100 each, with $b_1 \sim N(0, \sigma_1^2)$ for $\sigma_1^2 = 0.5^2$ and $b_2 \sim N(0, \sigma_2^2)$ for $\sigma_2^2 = 0.3^2$.

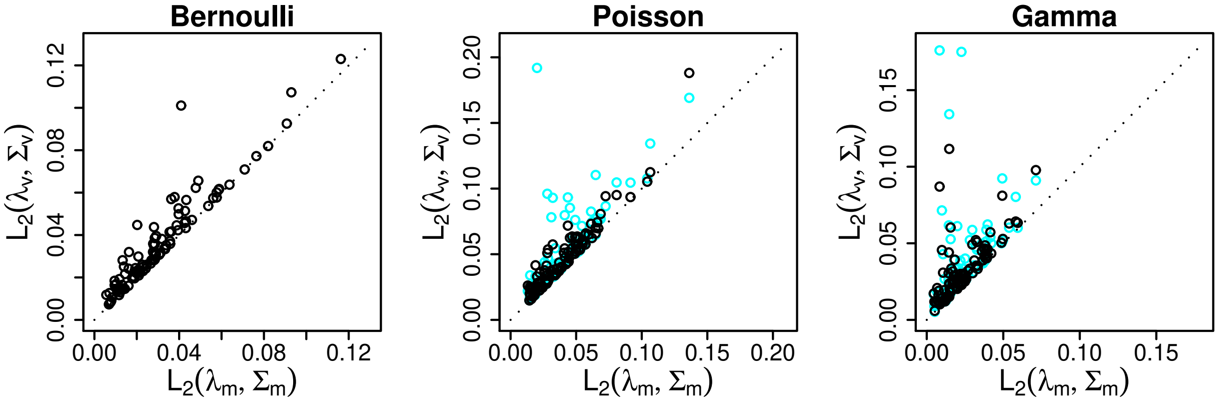

Figure 3: Simulations with Latent Random Effects. Left: Performance of (11) for Bernoulli data. Center: Performance of (12) for Poisson data. Right: Performance of (13) for gamma data. For Poisson and gamma data, results of $\alpha = 1$ are in faded circles and those of $\alpha = 1.4$ are in circles.

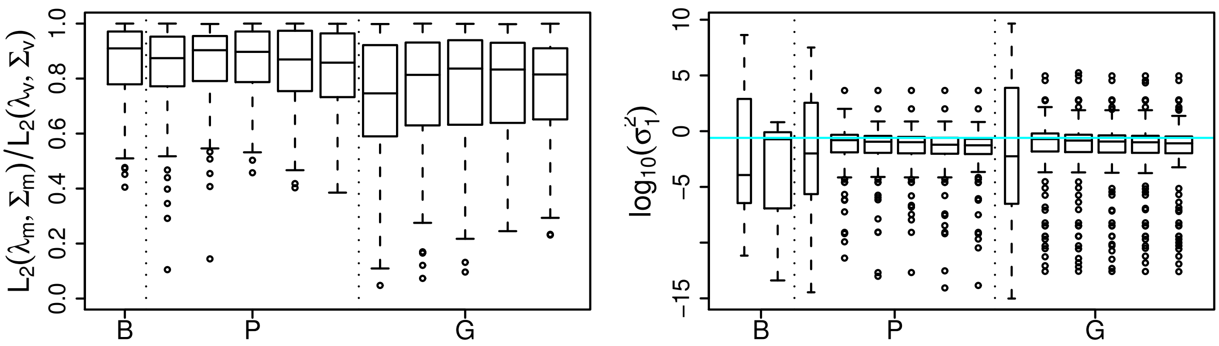

Figure 4: Simulation with Latent Random Effects. Left: Relative efficacy for Bernoulli (B), Poisson (P), and gamma (G) data; for P and G, the plots are for $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, in order. Right: Variance σ_1^2 "Estimated" through Σ_m (thinner boxes) and Σ_v (fatter boxes) for Bernoulli (B), Poisson (P), and gamma (G) data.

As in §5.1, one hundred replicates of samples were generated for each of the three distribution families, and cross-validated estimates were calculated. Estimates were also calculated with the tuning parameters (λ_m, Σ_m) minimizing the loss $L_2(\lambda, \Sigma)$ of (17). The counterparts of Figures 1 and 2 are given in Figures 3 and 4, but in the right frame of Figure 4, the box-plots are only for σ_1^2 through $a(\phi)/\sigma_1^2$ in Σ_m and Σ_v; the plots for σ_2^2 are similar. Note that the data only contain one sample each from $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$.

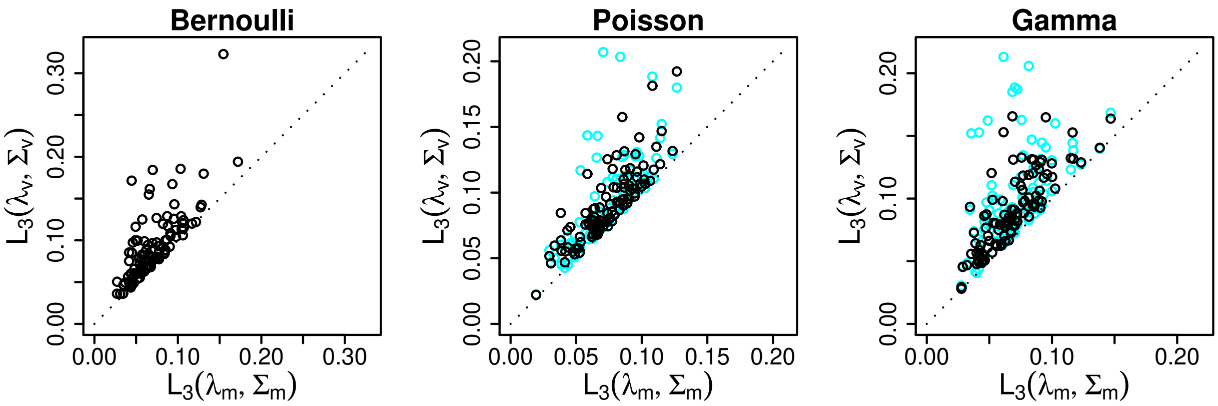

Figure 5: Simulations with Mixture Random Effects. Left: Performance of (11) for Bernoulli data. Center: Performance of (12) for Poisson data. Right: Performance of (13) for gamma data. For Poisson and gamma data, results of $\alpha = 1$ are in faded circles and those of $\alpha = 1.4$ are in circles.

5.4 Simulations with Mixture Random Effects

For mixture random effects, we add together the b_s of §5.2 and the b_c of §5.3, with the 20 subjects nested under the 2 clusters, 10 each. For each of the three distribution families, one hundred replicates of samples were generated, and estimates were calculated with (λ, Σ) minimizing the cross-validation scores (11), (12), or (13) in their respective settings. Also calculated are estimates minimizing the loss $L_3(\lambda, \Sigma)$ of (18). The counterpart of Figures 1 and 3 is shown in Firgure 5. The counterpart of Figures 2 and 4 look similar and is not shown here. The variance ratios $a(\phi)/\sigma_s^2$, $a(\phi)/\sigma_1^2$, and $a(\phi)/\sigma_2^2$ "estimated" through Σ_m and Σ_v remain highly unreliable.

6 Applications

We now report the analyses of a couple of real data sets using the techniques developed. Also reported are some timing results obtained on an Athlon MP2800+ workstation running FreeBSD 4.4 and R 1.6.2.

6.1 Treatment of Bacteriuria

Patients with acute spinal cord injury and bacteriuria (bacteria in urine) were randomly assigned to two treatment groups. Patients in the first group were treated for all episodes of urinary tract infection, whereas those in the second group were treated only if two specific symptoms occurred. Weekly binary indicator of bacteriuria was recorded for every patient over 4 to 16 weeks. A total of 72 patients were represented in the data, with 36 each in the two treatment groups. The data are listed in Joe (1997, §11.4), where further details and references can be found.

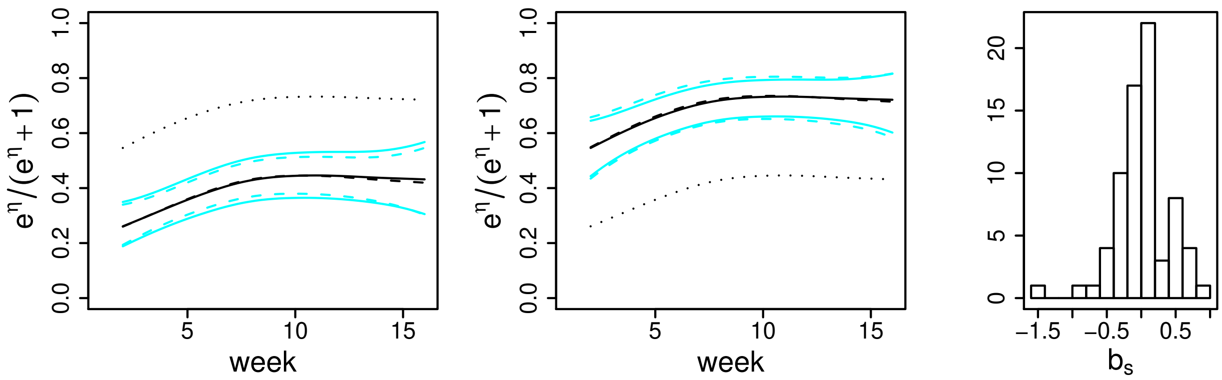

Figure 6: Cross-Validated Additive Cubic Spline Fit of Bacteriuria Data. Left: $\eta(t,1)$ with 95% Bayesian confidence intervals, in solid lines; $\eta(t,2)$ is in dotted line. Center: $\eta(t,2)$ with 95% Bayesian confidence intervals, in solid lines; $\eta(t,1)$ is in dotted line. Right: Histogram of $\hat{b}_s$. The dashed lines are from fit with separate σ_s^2 under different treatments.

Let $P(t,\tau,s)$ be the probability of bacteriuria of subject s at time t under treatment τ. We shall fit a logistic model of the form

$$\log\frac{P(t_i,\tau_i,s_i)}{1-P(t_i,\tau_i,s_i)} = \eta(t_i,\tau_i) + b_{s_i},$$

where $s_i \in \{1,\dots,72\}$. We use the tensor-product spline of Example 7 for $\eta(t_i,\tau_i)$ with $x = (x_{\langle 1\rangle}, x_{\langle 2\rangle}) = (t,\tau)$, where t is mapped into $[0,1]$ and $\tau \in \{1,2\}$. There are a total of 892 observations, but the week-1 bacteriuria indicator was positive for all patients. After removing the week-1 data, we have a sample size $n = 820$, and since there are only 30 distinctive x_i's (15 time points by 2 treatment levels), the "exact" solution can be computed with $q = 30$ non-random v_j's in (5).

The model was fitted with the tuning parameters $(\lambda/\theta_1, \lambda/\theta_{1,2}, \Sigma)$ selected by the cross-validation score of (11). Using the diagnostic tool of Gu (2003) based on Kullback-Leibler projection, the interaction $\eta_{1,2}(t,\tau)$ appeared negligible. Eliminating the interaction, an additive model $\eta(t_i,\tau_i) = \eta_\emptyset + \eta_1(t) + \eta_2(\tau)$ was fitted with the cross-validated (λ_v, Σ_v); $\eta(t,1)$ and $\eta(t,2)$ are parallel cubic splines in an additive model. Plotted in the left and center frames of Figure 6 are the estimated additive $\eta(t,\tau)$ on the probability scale along with the Bayesian confidence intervals, in solid lines; see the last paragraphs of §2 and §3 for brief discussions of the Bayesian confidence intervals and further references. Shown in the right frame is a histogram of the estimated $\hat{b}_s$. The sample variance of $\hat{b}_s$ was $s_b^2 = 0.1374$, while the σ_s^2 "estimated" through Σ_v was 0.3193; remember that the latter can be grossly misleading, as was shown in the simulations of §5.

Checking the estimated $\hat{b}_s$'s under the two treatments separately, it was revealed that those from $\tau = 1$ had less than half of the scatter when compared to those from $\tau = 2$. It was tempting to refit the models with two separate σ_s^2's assigned to b_s's under the two treatments, which we did. The interaction was again negligible. The fitted additive $\eta(t, \tau)$ largely remained the same, but the Bayesian confidence intervals were slightly tighter for $\eta(t, 1)$ and slightly wider for $\eta(t, 2)$; these are superimposed in Figure 6 in dashed lines. For $\tau = 1$, b_s were effectively suppressed, with $s_b^2 = 5.467 \times 10^{-11}$ and $\sigma_s^2 = 5.450 \times 10^{-6}$ through Σ_v; for $\tau = 2$, $s_b^2 = 0.4482$ and $\sigma_s^2 = 0.6452$ through Σ_v.

From the above analysis, it seems clear that the first treatment was superior to the second. The rate of bacteriuria increased steadily up to the 9th/10th week, then remained level. The responses from subjects under the first treatment appeared rather homogeneous, but the inter-subject variability under the second treatment seemed real and substantial.

The fits with interaction took about 36 and 74 CPU seconds, respectively, for common and separate σ_s^2; the respective timing results for the additive fits were about 7 and 17 CPU seconds.

6.2 Treatment of Epileptic Seizure

Patients suffering from simple or complex partial seizures were randomized to receive either the antiepileptic drug progabide or a placebo, as an adjuvant to standard chemotherapy. The patients were followed up for 8 weeks in 4 biweekly clinic visits and the biweekly seizure counts were collected. Also collected were the baseline seizure counts over the 8 weeks prior to the trial and the age of the patients. A total of 59 patients were represented in the data, with 31 receiving progabide and 28 receiving placebo. The data are listed in Thall and Vail (1990), where further details can be found.

Let $\Lambda(x, s)$ be the seizure intensity of subject s at covariate x, where $x = (x_{\langle 1 \rangle}, x_{\langle 2 \rangle}, x_{\langle 3 \rangle}, x_{\langle 4 \rangle})$ consists of the treatment (2 levels), the time of clinic visit (4 points), the baseline seizure count, and the age of patient, in order. We shall fit an Poisson model of the form

$$\log \Lambda(x_i, s_i) = \eta_\emptyset + \eta_1(x_{i\langle 1 \rangle}) + \eta_2(x_{i\langle 2 \rangle}) + \eta_3(x_{i\langle 3 \rangle}) + \eta_4(x_{i\langle 4 \rangle}) + b_{s_i}$$

where $s_i \in \{1, \ldots, 59\}$; η_1 is plus or minus a constant depending on the treatment level, and η_2, η_3, η_4 are additive cubic splines with

$$J(\eta) = \theta_2^{-1} \int \ddot{\eta}_2^2 dx_{\langle 2 \rangle} + \theta_3^{-1} \int \ddot{\eta}_3^2 dx_{\langle 3 \rangle} + \theta_4^{-1} \int \ddot{\eta}_4^2 dx_{\langle 4 \rangle}.$$

The sample size is $n = 59 \times 4 = 236$ and the "exact" solution is available with $q = 59$ non-random v_j's in (5). A log transform is applied to the baseline seizure counts to spread out the data more evenly.

The model was fitted with the tuning parameters $(\lambda/\theta_2, \lambda/\theta_3, \lambda/\theta_4, \Sigma)$ selected by the cross-validation score (12) with $\alpha = 1.4$. Using the diagnostic tool of Gu (2003), $\eta_1 + \eta_2 + \eta_4$ appeared

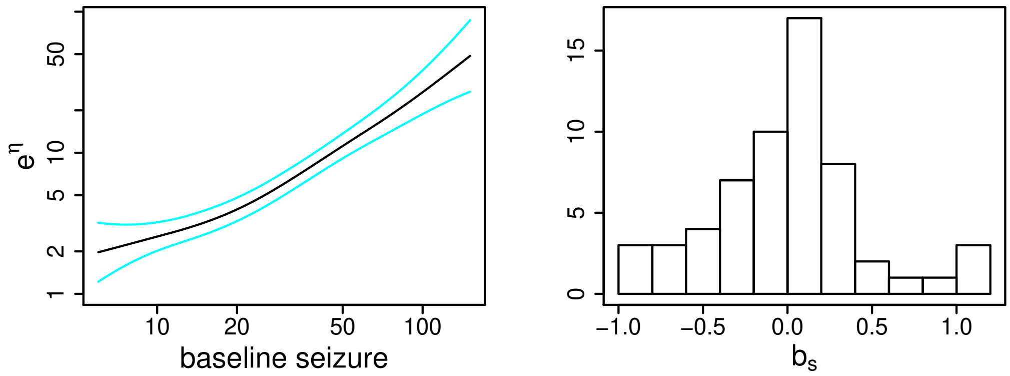

Figure 7: Cross-Validated Cubic Spline Fit of Seizure Count Data. Left: η with 95% Bayesian confidence intervals. Right: Histogram of $\hat{b}_s$.

negligible, so we dropped the terms and fitted a cubic spline model of Example 6 with the baseline seizure count as the sole covariate. The final fit is plotted in Figure 7. The sample variance of $\hat{b}_s$ was $s_b^2 = 0.1916$ and the σ_s^2 "estimated" through Σ_v was 0.1966.

The analysis shows that the treatment makes little difference, nor do the time of clinic visit and the age of patient. The baseline seizure count seems to be the dominant factor, and the inter-subject variability appears appreciable.

The fit with four covariates took about 22 CPU seconds and the final fit with one covariate took about 6 CPU seconds.

7 Discussion

In this article, we have studied an approach to the estimation of nonparametric mixed-effect models in logistic regression, Poisson regression, and gamma regression. Efficient computational strategies have been proposed through the use of existing algorithms for independent data, and the effective selection of tuning parameters by certain cross-validation techniques has been empirically evaluated in terms of a few loss functions. Practical applications of the techniques are also illustrated through the analyses of a couple of real data sets.

It should be noted that more simulations were done than what have been reported in §5, with pretty much the same results, qualitatively. Also note the single cross-validation scores for use in each of the distribution families that target the different loss functions $L_1(\lambda, \Sigma)$, $L_2(\lambda, \Sigma)$, or $L_3(\lambda, \Sigma)$, whichever is appropriate in the application settings; parallel results in Gaussian regression can be found in Gu and Ma (2003), with both theoretical proofs and empirical verifications. Theoretical justifications of the cross-validation scores of §4 are not available even in the absence of random effects, however. The theory of Gu and Ma (2003) may apply if the indirect cross-validation of Gu (1992a) is employed, which however is numerically less efficient; see Gu and Wang (2003, §3) for discussions concerning the comparison of direct and indirect cross-validations.

The key to the structural simplicity and computational efficiency of our approach is the inclusion of the latent variables **b** in the estimation, which turns the "variance components" into "mean components." Similar use of the strategy in parametric estimation can be found in Therneau and Grambsch (1998).

The calculations reported in this article were performed in R (Ihaka and Gentleman 1996), an open-source clone of S/S-PLUS. The code will be made available to the public in the near future, after further polishing of the user-interface.

Acknowledgements

The work of Ping Ma was done while he was a graduate student at Purdue University. This research was supported by National Institutes of Health under Grant R33HL68515.

References

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc. 88*, 9–25.

Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: SIAM. Corrected reprint of the 1983 original.

Gu, C. (1992a). Cross-validating non-Gaussian data. *J. Comput. Graph. Statist. 1*, 169–179.

Gu, C. (1992b). Penalized likelihood regression: A Bayesian analysis. *Statist. Sin. 2*, 255–264.

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

Gu, C. (2003). Model diagnostics for smoothing spline ANOVA models. Technical report, Department of Statistics, Purdue University, West Lafayette, IN.

Gu, C. and Y.-J. Kim (2002). Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist. 30*, 619–628.

Gu, C. and P. Ma (2003). Optimal smoothing in nonparametric mixed-effect models. Revised for *Ann. Statist.*

Gu, C. and J. Wang (2003). Penalized likelihood density estimation: Direct cross validation and scalable approximation. *Statist. Sin. 13*, 811–826.

Gu, C. and D. Xiang (2001). Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited. *J. Comput. Graph. Statist. 10*, 581–591.

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist. 5*, 299–314.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Karcher, P. and Y. Wang (2001). Generalized nonparametric mixed effects models. *J. Comput. Graph. Statist. 10*, 641–655.

Kim, Y.-J. (2003). *Smoothing Spline Regression: Scalable Computation and Cross-Validation*. Ph. D. thesis, Purdue University, West Lafayette, IN.

Kim, Y.-J. and C. Gu (2003). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. Revised for *J. Roy. Statist. Soc. Ser. B*.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Statist. Soc. Ser. B 61*, 381–400.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc. 92*, 162–170.

Thall, P. F. and S. C. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics 46*, 657–671.

Therneau, T. M. and P. M. Grambsch (1998). Penalized Cox models and frailty. Technical report, Division of Biostatistics, Mayo Clinic, Rochester, MN.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B 40*, 364–372.

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

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sin. 6*, 675–692.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. Amer. Statist. Assoc. 86*, 79–86.