J. R. Statist. Soc. B (1999)
61, *Part* 2, *pp.* 381–400

Inference in generalized additive mixed models by using smoothing splines

Xihong Lin

University of Michigan, Ann Arbor, USA

and Daowen Zhang

North Carolina State University, Raleigh, USA

[Received October 1997. Revised July 1998]

Summary. Generalized additive mixed models are proposed for overdispersed and correlated data, which arise frequently in studies involving clustered, hierarchical and spatial designs. This class of models allows flexible functional dependence of an outcome variable on covariates by using nonparametric regression, while accounting for correlation between observations by using random effects. We estimate nonparametric functions by using smoothing splines and jointly estimate smoothing parameters and variance components by using marginal quasi-likelihood. Because numerical integration is often required by maximizing the objective functions, double penalized quasi-likelihood is proposed to make approximate inference. Frequentist and Bayesian inferences are compared. A key feature of the method proposed is that it allows us to make systematic inference on all model components within a unified parametric mixed model framework and can be easily implemented by fitting a working generalized linear mixed model by using existing statistical software. A bias correction procedure is also proposed to improve the performance of double penalized quasi-likelihood for sparse data. We illustrate the method with an application to infectious disease data and we evaluate its performance through simulation.

Keywords: Correlated data; Generalized linear mixed models; Laplace approximation; Marginal quasi-likelihood; Nonparametric regression; Penalized quasi-likelihood; Smoothing parameters; Variance components

1. Introduction

Generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993) provide a unified likelihood framework for parametric regression of a variety of overdispersed and correlated outcomes. Data of this type arise in many fields of research, such as longitudinal studies, survey sampling, clinical trials and disease mapping. A major difficulty in making inference in GLMMs is that a full likelihood analysis is burdened by often intractable numerical integration. Various approximate inference procedures (Breslow and Clayton, 1993; Lee and Nelder, 1996; Lin and Breslow, 1996) and Bayesian procedures using EM algorithms and Gibbs sampling (McCulloch, 1997; Zeger and Karim, 1991) have been proposed. For discussion on full maximum likelihood estimation, see Aitkin (1998).

A key feature of GLMMs is that they use a parametric mean function to model covariate effects, while accommodating overdispersion and correlation by adding random effects to the

Address for correspondence: Xihong Lin, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA.
E-mail: xlin@sph.umich.edu

 1369–7412/99/61381

linear predictor. However, this parametric mean assumption may not always be desirable, since appropriate functional forms of the covariates may not be known in advance and the outcome variable may depend on the covariates in a complicated manner. It is hence of substantial interest to develop a nonparametric regression model for correlated data by incorporating a nonparametric mean function in GLMMs. This will allow more flexible functional dependence of the outcome variable on the covariates.

There are very many references on nonparametric regression with independent data using kernel and spline methods (Härdle, 1990; Green and Silverman, 1994). The generalized additive models of Hastie and Tibshirani (1990) are widely used and well understood. However, only very limited work has been done on nonparametric regression when the data are correlated. Most researchers have restricted their attention to longitudinal data with normally distributed outcomes and a single nonparametric function (Hart, 1991; Rice and Silverman, 1991). Several researchers have incorporated a nonparametric time function in linear mixed models (Zeger and Diggle, 1994; Zhang *et al.*, 1998; Verbyla *et al.*, 1998). For non-Gaussian longitudinal data, Wild and Yee (1996) and Berhane and Tibshirani (1998) extended generalized additive models to generalized estimating equations (Liang and Zeger, 1986). There are not many references on modelling correlated non-Gaussian outcomes nonparametrically within the mixed effects model framework. See Verbyla (1995) for discussion on mixed model formulation of smoothing splines in generalized linear models for independent non-Gaussian data.

Nonparametric regression with correlated data faces many new challenges. In addition to developing an inference procedure for nonparametric functions, we also need to consider how to draw inference on correlation parameters. Another critical issue, whose importance has been emphasized by many (Green and Silverman, 1994; Wahba, 1978), is how to select good estimators of smoothing parameters and bandwidth parameters. Very limited work has been done on these issues, especially estimation of the correlation parameters and the smoothing parameters. Conventional data-driven methods for smoothing parameter estimation are challenged with new problems. For example, although cross-validation (Rice and Silverman, 1991) is a reasonable approach to selecting the smoothing parameters for clustered data, it is often computationally expensive and subsequent inference on the correlation parameters is difficult (Zeger and Diggle, 1994) and it fails for crossed designs and spatial data. It is hence of substantial interest to develop a systematic procedure to make inference on all model parameters.

In this paper, we propose generalized additive mixed models (GAMMs), which are an additive extension of GLMMs in the spirit of Hastie and Tibshirani (1990). This new class of models uses additive nonparametric functions to model covariate effects while accounting for overdispersion and correlation by adding random effects to the additive predictor. GAMMs encompass nested and crossed designs and are applicable to clustered, hierarchical and spatial data.

We estimate the nonparametric functions by using smoothing splines and jointly estimate the smoothing parameters and the variance components by using marginal quasi-likelihood. This marginal quasi-likelihood approach is an extension of the restricted maximum likelihood (REML) approach used by Wahba (1985) and Kohn *et al.* (1991) in the classical nonparametric regression model (Kohn *et al.* (1991), equation (2.1)), and by Zhang *et al.* (1998), Brumback and Rice (1998) and Wang (1998) in Gaussian nonparametric mixed models, where they treated the smoothing parameter as an extra variance component. Because numerical integration is often required by maximizing the objective functions, double penalized quasi-likelihood (DPQL) is proposed to make approximate inference.

Frequentist and Bayesian inferences are compared. A key feature of the method proposed is that it allows us to make systematic inference on all model components of GAMMs within a unified parametric mixed model framework. Specifically, our estimation of the nonparametric functions, the smoothing parameters and the variance components in GAMMs can proceed by fitting a working GLMM using existing statistical software, which iteratively fits a linear mixed model to a modified dependent variable. When the data are sparse (e.g. binary), the DPQL estimators of the variance components are found to be subject to considerable bias. A bias correction procedure is hence proposed to improve its performance. We illustrate the method with an application to infectious disease data and we evaluate its performance through simulation.

2. The generalized additive mixed model

Suppose that observations of the ith of n units consist of an outcome variable y_i and p covariates $x_i = (1, x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ associated with fixed effects and a $q \times 1$ vector of covariates z_i associated with random effects. Given a $q \times 1$ vector b of random effects, the observations y_i are assumed to be conditionally independent with means $E(y_i|b) = \mu_i^b$ and variances $\mathrm{var}(y_i|b) = \phi m_i^{-1} v(\mu_i^b)$, where $v(\cdot)$ is a specified variance function, m_i is a prior weight (e.g. a binomial denominator) and ϕ is a scale parameter, and follow a generalized additive model

$$g(\mu_i^b) = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + z_i^{\mathrm{T}} b, \tag{1}$$

where $g(\cdot)$ is a monotonic differentiable link function, $f_j(\cdot)$ is a centred twice-differentiable smooth function, the random effects b are assumed to be distributed as $N\{0, D(\theta)\}$ and θ is a $c \times 1$ vector of variance components.

A key feature of the GAMM (1) is that additive nonparametric functions are used to model covariate effects and random effects are used to model correlation between observations. If $f_j(\cdot)$ is a linear function, the GAMM (1) reduces to the GLMM of Breslow and Clayton (1993). Zeger and Diggle (1994) and Zhang *et al.* (1998) considered a special case of the GAMM (1), a semiparametric mixed model, where they assumed a single nonparametric time function $f(\cdot)$ and longitudinal data with normally distributed outcomes.

Model formulation (1) encompasses various study designs, such as clustered, hierarchical and spatial designs. This is because we can specify a flexible covariance structure of the random effects b. For example, for longitudinal data, the random effects b can be decomposed into a random intercept and a stochastic process (Zeger and Diggle, 1994; Zhang *et al.*, 1998). For hierarchical (multilevel) data, they can be partitioned to represent different levels of a hierarchy, e.g. a centre, physician and patient in a multicentre clinical trial (Lin and Breslow, 1996). For spatial data, which are common in disease mapping and ecological studies, they can be used to model spatial correlation, which is often assumed as a function of the Euclidean distance between every two regions (Cressie, 1993), or a constant between every two adjacent regions (Breslow and Clayton, 1993).

The integrated log-quasi-likelihood of $\{\beta_0, f_1(\cdot), \ldots, f_p(\cdot), \theta\}$ is (compare Breslow and Clayton (1993), equation (2))

$$\exp[l\{y; \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \theta\}] \propto |D|^{-1/2} \int \exp\left\{ -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i; \mu_i^b) - \frac{1}{2} b^{\mathrm{T}} D^{-1} b \right\} \mathrm{d}b, \tag{2}$$

where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and

$$d_i(y_i; \mu_i^b) \propto -2\int_{y_i}^{\mu_i^b} m_i(y_i - u)/v(u)\,\mathrm{d}u$$

defines the conditional deviance function of $\{\beta_0, f_1(\cdot), \ldots, f_p(\cdot)\}$ given b. For simplicity, we here assume that D is a full rank matrix. If not, the Moore–Penrose generalized inverse may be used.

Statistical inference in the GAMM (1) involves inference on the nonparametric functions $f_j(\cdot)$, which often requires the estimation of smoothing parameters, say λ, and inference on the variance components θ. In the next two sections, we shall first discuss how to construct natural cubic smoothing spline estimators of the $f_j(\cdot)$ when λ and θ are known; then we propose to estimate λ and θ jointly by using marginal quasi-likelihood.

3. Inference on the nonparametric functions

3.1. Natural cubic smoothing spline estimation

Since the $f_j(\cdot)$ are infinite dimensional unknown parameters, we consider estimating them by using natural cubic smoothing splines. Using the results of O'Sullivan *et al.* (1986) and noting that equation (2) is a continuous linear *functional* of the $f_j(\cdot)$, one can show that, for given values of λ and θ, the natural cubic smoothing spline estimators of the $f_j(\cdot)$ maximize the penalized log-quasi-likelihood

$$l\{y; \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \theta\} - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \int_{s_j}^{t_j} f_j''(x)^2\,\mathrm{d}x = l(y; \beta_0, f_1, \ldots, f_p, \theta) - \frac{1}{2}\sum_{j=1}^{p}\lambda_j f_j^{\mathrm{T}} K_j f_j, \tag{3}$$

where (s_j, t_j) defines the range of the jth covariate and $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ is a vector of smoothing parameters and controls the trade-off between the goodness of fit and the smoothness of the estimated functions. Here f_j is an $r_j \times 1$ unknown vector of the values of $f_j(\cdot)$ evaluated at the r_j ordered *distinct* values of the x_{ij} $(i = 1, \ldots, n)$, and K_j is the corresponding non-negative definite smoothing matrix (Green and Silverman (1994), equation (2.3)).

In matrix notation, with $\mu^b = (\mu_1^b, \ldots, \mu_n^b)^{\mathrm{T}}$, $g(\mu^b) = \{g(\mu_1^b, \ldots, g(\mu_n^b)\}^{\mathrm{T}}$ and $Z = (z_1, \ldots, z_n)^{\mathrm{T}}$, GAMM (1) can be written

$$g(\mu^b) = \mathbf{1}\beta_0 + N_1 f_1 + \ldots + N_p f_p + Zb, \tag{4}$$

where $\mathbf{1}$ is an $n \times 1$ vector of 1s and N_j is an $n \times r_j$ incidence matrix defined in a way similar to that given in Green and Silverman (1994), section 4.3.1, such that the ith component of $N_j f_j$ is $f_j(x_{ij})$.

Since the evaluation of $l(y; \beta_0, f_1, \ldots, f_p, \theta)$ in expression (2) requires numerical integration, except for the Gaussian case, it is often difficult to calculate full natural cubic smoothing spline estimators of the f_j by directly maximizing expression (3). An approximation is hence proposed in the next section. For discussion on full cubic smoothing spline calculations by using Monte Carlo simulation, see Section 8.

3.2. Double penalized quasi-likelihood

In view of often intractable numerical integration required for maximizing expression (3), we approximate it by applying the Laplace method to $l(y; \beta_0, f_1, \ldots, f_p, \theta)$ in expression (2)

(Tierney and Kadane, 1986). Maximizing the resultant approximation to equation (3) yields approximate natural cubic spline estimators of the f_j. Note that this approximation is exact for normally distributed outcomes with identity link function. Specifically, we approximate expression (2) by making a quadratic expansion of the exponent of the integrand about its maximum point before integration. Ignoring the first determinant term of the resultant Laplace approximation (Breslow and Clayton, 1993), some calculation shows that approximate natural cubic smoothing spline estimators $(\hat{\beta}_0, \hat{f}_1, \ldots, \hat{f}_p)$ can be obtained by maximizing the following DPQL with respect to $(\beta_0, f_1, \ldots, f_p)$ and b:

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i; \mu_i^b) - \frac{1}{2} b^{\mathrm{T}} D^{-1} b - \frac{1}{2}\sum_{j=1}^{p} \lambda_j f_j^{\mathrm{T}} K_j f_j. \tag{5}$$

See Appendix A for a detailed derivation of expression (5). The first penalty term $b^{\mathrm{T}}D^{-1}b/2$ results from the Laplace approximation of expression (2), whereas the second penalty term $\lambda_j f_j^{\mathrm{T}} K_j f_j/2$ results from the natural cubic smoothing spline property of f_j.

Differentiating expression (5) with respect to $(\beta_0, f_1, \ldots, f_p)$ and b yields their estimating equations as

$$\begin{gathered} \mathbf{1}^{\mathrm{T}} W\Delta(y-\mu^b) = 0, \\ N_j^{\mathrm{T}} W\Delta(y-\mu^b) - \lambda_j K_j f_j = 0 \qquad (j = 1, \ldots, p), \\ Z^{\mathrm{T}} W\Delta(y-\mu^b) - D^{-1} b = 0, \end{gathered} \tag{6}$$

where $\Delta = \mathrm{diag}\{g'(\mu_i^b)\}$, $W = \mathrm{diag}[\{\phi m_i^{-1} v(\mu_i^b) g'(\mu_i^b)^2\}^{-1}]$ is a modified generalized additive model working weight matrix and f_j needs to satisfy $f_j^{\mathrm{T}}\mathbf{1} = 0$ so that f_j is centred. Equation (6) can be solved by using the Fisher scoring algorithm (compare Hastie and Tibshirani (1990), equation (6.16)),

$$\begin{pmatrix} \mathbf{1} & S_0 N_1 & \cdots & S_0 N_p & S_0 Z \\ S_1\mathbf{1} & I & \cdots & S_1 N_p & S_1 Z \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ S_p\mathbf{1} & S_p N_1 & \cdots & I & S_p Z \\ S_b\mathbf{1} & S_b N_1 & \cdots & S_b N_p & I \end{pmatrix} \begin{pmatrix} \beta_0 \\ f_1 \\ \vdots \\ f_p \\ b \end{pmatrix} = \begin{pmatrix} S_0 Y \\ S_1 Y \\ \vdots \\ S_p Y \\ S_b Y \end{pmatrix}, \tag{7}$$

where

$$Y = \beta_0\mathbf{1} + \sum_{j=1}^{p} N_j f_j + Zb + \Delta(y-\mu^b)$$

is a modified generalized additive model working vector, S_j is a centred smoother for f_j and satisfies $S_j^{\mathrm{T}}\mathbf{1} = 0$, and S_j and S_b are defined as

$$S_0 = (\mathbf{1}^{\mathrm{T}} W\mathbf{1})^{-1}\mathbf{1}^{\mathrm{T}} W,$$

$$S_j = \left(I - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{r_j}\right)(N_j^{\mathrm{T}} W N_j + \lambda_j K_j)^{-1} N_j^{\mathrm{T}} W \qquad (j = 1, \ldots, p),$$

$$S_b = (Z^{\mathrm{T}} W Z + D^{-1})^{-1} Z^{\mathrm{T}} W.$$

It follows that the resultant estimators $\hat{f}_j$ are centred.

If the design matrix $X = (x_1, \ldots, x_n)^T$ is of full rank, i.e. no concurvity (Hastie and Tibshirani, 1990), we can use the results in Section 3.3 to show that the solution to equation (7) is unique and the following back-fitting algorithm converges to this unique solution:

$$\beta_0 = S_0\left(Y - \sum_{j=1}^{p} N_j f_j - Zb\right),$$

$$f_j = S_j\left(Y - \beta_0 \mathbf{1} - \sum_{k \neq j} N_k f_k - Zb\right) \qquad (j = 1, \ldots, p),$$

$$b = S_b\left(Y - \beta_0 \mathbf{1} - \sum_{j=1}^{p} N_j f_j\right).$$

The last equation suggests that $\hat{b}$ can be viewed as an empirical Bayes estimator.

3.3. The generalized linear mixed model representation

We show in this section that the DPQL estimators $\hat{f}_j$ defined in Section 3.2 can be easily obtained by fitting a GLMM using existing statistical software. This GLMM representation provides a foundation for our joint estimation procedure for the smoothing parameters λ and the variance components θ in Section 4.

Following Green (1987), equation (4.2), and Zhang *et al.* (1998), equation (10), and noting that f_j is a centred parameter vector, we can reparameterize f_j in terms of β_j (scalar) and a_j ($(r_j - 2) \times 1$) via a one-to-one transformation as

$$f_j = X_j \beta_j + B_j a_j, \tag{8}$$

where X_j is an $r_j \times 1$ vector containing the r_j centred ordered *distinct* values of the x_{ij} ($i = 1, \ldots, n$), and $B_j = L_j (L_j^T L_j)^{-1}$ and L_j is an $r_j \times (r_j - 2)$ full rank matrix satisfying $K_j = L_j L_j^T$ and $L_j^T X_j = 0$. Using the identity $f_j^T K_j f_j = a_j^T a_j$, DPQL (5) becomes (compare Breslow and Clayton (1993), equation (6), and Zhang *et al.* (1998), equation (11))

$$-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y; \mu_i^b) - \frac{1}{2} b^T D^{-1} b - \frac{1}{2} a^T \Lambda^{-1} a, \tag{9}$$

where $a = (a_1^T, \ldots, a_p^T)^T$ and $\Lambda = \mathrm{diag}(\tau_1 I, \ldots, \tau_p I)$ with $\tau_j = 1/\lambda_j$. A small value of $\tau = (\tau_1, \ldots, \tau_p)^T$ corresponds to oversmoothing.

Plugging equation (8) into equation (4), equation (9) suggests that, given θ and τ, the DPQL estimators $\hat{f}_j$ can be obtained by fitting the following GLMM by using Breslow and Clayton's (1993) penalized quasi-likelihood approach:

$$g(\mu^b) = X\beta + Ba + Zb, \tag{10}$$

where X was defined at the end of Section 3.2 and is also equal to $X = (\mathbf{1}, N_1 X_1, \ldots, N_p X_p)$, $B = (N_1 B_1, \ldots, N_p B_p)$, $\beta = (\beta_0, \ldots, \beta_p)^T$ is a $(p+1) \times 1$ vector of regression coefficients and a and b are independent random effects with distributions $a \sim N(0, \Lambda)$ and $b \sim N(0, D)$. The DPQL estimator $\hat{f}_j$ is calculated as $\hat{f}_j = X_j \hat{\beta}_j + B_j \hat{a}_j$, which is a linear combination of the Breslow and Clayton (1993) penalized quasi-likelihood estimators of the fixed effect $\hat{\beta}_j$ and the random effects $\hat{a}_j$ in the working GLMM (10) and can be obtained by fitting the working GLMM (10) by using existing statistical software, such as the SAS macro `GLIMMIX` (Wolfinger, 1996).

Specifically, the maximization of expression (9) with respect to (β, a, b) can proceed by using the Fisher scoring algorithm to solve

$$\begin{pmatrix} X^{\mathrm{T}}WX & X^{\mathrm{T}}WB & X^{\mathrm{T}}WZ \\ B^{\mathrm{T}}WX & B^{\mathrm{T}}WB+\Lambda^{-1} & B^{\mathrm{T}}WZ \\ Z^{\mathrm{T}}WX & Z^{\mathrm{T}}WB & Z^{\mathrm{T}}WZ+D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \\ b \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}WY \\ B^{\mathrm{T}}WY \\ Z^{\mathrm{T}}WY \end{pmatrix}, \tag{11}$$

where Y is the working vector defined in Section 3.2. One can easily show that equation (11) has a unique solution $\hat{f}_j = X_j\hat{\beta}_j + B_j\hat{a}_j$ $(j = 1, \ldots, p)$ if X is of full rank, and the $\hat{f}_j$ obtained from equation (11) are identical with those obtained from equation (7).

An examination of equation (11) shows that it corresponds to the normal equation of the best linear unbiased predictors (BLUPs) of β and (a, b) under the linear mixed model

$$Y = X\beta + Ba + Zb + \epsilon, \tag{12}$$

where a and b are independent random effects with $a \sim N(0, \Lambda)$ and $b \sim N(0, D)$ and $\epsilon \sim N(0, W^{-1})$. This suggests that the DPQL estimators $\hat{f}_j$ and the random effect estimators $\hat{b}$ can be easily obtained using the BLUPs by iteratively fitting model (12) to the working vector Y.

To compute the covariance matrix of $\hat{f}_j$, it is more convenient to calculate β and a by using

$$\begin{pmatrix} X^{\mathrm{T}}R^{-1}X & X^{\mathrm{T}}R^{-1}B \\ B^{\mathrm{T}}R^{-1}X & B^{\mathrm{T}}R^{-1}B+\Lambda^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}R^{-1}Y \\ B^{\mathrm{T}}R^{-1}Y \end{pmatrix}, \tag{13}$$

where $R = W^{-1} + ZDZ^{\mathrm{T}}$. Denoting by H the coefficient matrix on the left-hand side of equation (13) and $H_0 = (X, B)^{\mathrm{T}}R^{-1}(X, B)$, the approximate covariance matrix of $\hat{\beta}$ and $\hat{a}$ is

$$\mathrm{cov}(\hat{\beta}, \hat{a}) = H^{-1}H_0H^{-1}. \tag{14}$$

It follows that the approximate covariance matrix of $\hat{f}_j$ is $(X_j, B_j)\,\mathrm{cov}(\hat{\beta}_j, \hat{a}_j)(X_j, B_j)^{\mathrm{T}}$, where $\mathrm{cov}(\hat{\beta}_j, \hat{a}_j)$ can be easily obtained from the corresponding blocks of $H^{-1}H_0H^{-1}$. Here we assume that the $f_j(\cdot)$ are fixed smooth functions in calculating the covariances of the $\hat{f}_j$.

3.4. *Bayesian formulation and inference*

We study in this section how to derive from a Bayesian perspective the natural cubic smoothing spline estimators of the f_j discussed in Sections 3.1–3.3, and how to derive the Bayesian standard errors of the $\hat{f}_j$ in the same spirit as Wahba (1983) and Zhang *et al.* (1998), as an alternative to the frequentist standard errors calculated at the end of Section 3.3. These standard errors will be used to construct confidence intervals for the f_j and their performance will be compared.

We first consider deriving the natural cubic smoothing spline estimators of the f_j within a Bayesian framework. Assuming that f_j is centred and has a prior log-density whose kernel is $-\lambda_j f_j^{\mathrm{T}} K_j f_j/2$, it can be easily seen from equation (3) that the full natural cubic smoothing spline estimators of the f_j are the posterior modes of the integrated quasi-likelihood function $l(y; \beta_0, f_1, \ldots, f_p, \theta)$. The DPQL estimators $\hat{f}_j$ are an approximation of these posterior modes. Alternatively, following Wahba (1978), the posterior mode property of $\hat{f}_j$ can be obtained by assuming a centred partially improper integrated Wiener prior for f_j. The partially integrated Wiener prior specification takes the same form as equation (8) except that

$$B_j = (I - \mathbf{1}\mathbf{1}^{\mathrm{T}}/r_j)\Sigma_j^{1/2},$$

where Σ_j is the $r_j \times r_j$ covariance matrix of an integrated Wiener process evaluated at X_j, and a_j has a normal prior $N(0, \tau_j I)$ and β_j has a flat prior. This partially integrated Wiener formulation provides an alternative GLMM representation, which takes the same form as equation (10) except that the design matrix B_j is replaced by $(I - \mathbf{1}\mathbf{1}^{\mathrm{T}}/r_j)\Sigma_j^{1/2}$.

This Bayesian smoothing spline formulation motivates us to consider calculating the Bayesian standard error of $\hat{f}_j$ similar to those given in Wahba (1983) and Zhang *et al.* (1998). To illustrate the idea, first consider the classical nonparametric regression model

$$y_i = f(x_i) + \epsilon_i, \tag{15}$$

where the ϵ_i are independent random errors following $N(0, \sigma^2)$. Wahba (1983) suggested to estimate the covariance of the natural cubic smoothing spline estimator $\hat{f}$ using its posterior covariance with the prior given above. Specifically, let $A = (I + \sigma^2\lambda K)^{-1}$, where K is the smoothing matrix. The Bayesian covariance matrix of $\hat{f}$ is $\sigma^2 A$, whereas its frequentist counterpart is $\sigma^2 A^2$ and is calculated by assuming the true $f(\cdot)$ to be a fixed smooth function (Hastie and Tibshirani (1990), section 3.8.1). As Wahba (1978) noted, in contrast with the frequentist standard errors, the Bayesian standard errors of $\hat{f}$ account for the bias in $\hat{f}$. Wahba (1983) showed that the confidence intervals of f calculated using the Bayesian standard errors have good coverage probabilities when the true $f(x)$ is a fixed smooth function.

Zhang *et al.* (1998) extended Wahba's (1983) Bayesian confidence interval calculations to estimate the confidence intervals of the nonparametric time function in a semiparametric linear mixed model for longitudinal Gaussian data. As indicated in Section 2, their model can be regarded as a special case of the GAMM (1). The simulation study results of Zhang *et al.* (1998) show that, when the true time function is a fixed smooth function, the coverage probabilities of the Bayesian confidence intervals of f are comparable with and sometimes slightly better than the frequentist counterparts.

It is hence of potential interest to derive the Bayesian standard errors of the DPQL estimators $\hat{f}_j$ under the GAMM (1), and to compare their performance with the frequentist standard errors, which were calculated at the end of Section 3.3 under the assumption that the $f_j(\cdot)$ are fixed smooth functions. Assuming that the f_j have priors given above, some calculation shows that the approximate Bayesian covariance matrix of the DPQL estimator of $(\hat{\beta}, \hat{a})$ is

$$\mathrm{cov}_{\mathrm{B}}(\hat{\beta}, \hat{a}) = H^{-1}, \tag{16}$$

which has a simpler form than its frequentist counterpart in equation (14). It follows that the approximate Bayesian covariance of $\hat{f}_j$ is $(X_j, B_j)\,\mathrm{cov}_{\mathrm{B}}(\hat{\beta}_j, \hat{a}_j)(X_j, B_j)^{\mathrm{T}}$, where $\mathrm{cov}_{\mathrm{B}}(\hat{\beta}_j, \hat{a}_j)$ can be easily obtained from the corresponding blocks of H^{-1}. We compare through simulation in Section 7 the coverage probabilities of the frequentist and Bayesian confidence intervals of the f_j in the situation of our primary interest, i.e. when the true $f_j(\cdot)$ are fixed smooth functions.

4. Inference on the smoothing parameters and the variance components

We assumed in Section 3 that the smoothing parameters λ and the variance components θ are known when we make inference on the nonparametric functions f_j. However, they often need to be estimated from the data. In this section, we propose to estimate λ and θ jointly by using marginal quasi-likelihood by extending the REML approach of Wahba (1985), Kohn *et al.*

(1991) and Zhang *et al.* (1998). A key feature of this approach is that λ and θ can be easily obtained by fitting the working GLMM (10) via iteratively fitting the working linear mixed model (12) using REML with $\tau = (1/\lambda_1, \ldots, 1/\lambda_p)^T$ treated as extra variance components in addition to θ.

4.1. Motivation of marginal quasi-likelihood: review of restricted maximum likelihood

In this section, we shall briefly review the use of REML to estimate the smoothing parameters λ and the variance components θ when outcomes are normally distributed. This will provide a motivation for our joint estimation procedure for λ and θ by using marginal quasi-likelihood discussed in the next section.

Under classical nonparametric regression model (15), Wahba (1985) and Kohn *et al.* (1991) proposed to estimate the smoothing parameter λ by maximizing a marginal likelihood. The marginal likelihood of $\tau = 1/\lambda$ is constructed by assuming that $f(x)$ has a prior specified in Section 3.4 in the form of equation (8) with $a \sim N(0, \tau I)$ and a flat prior for β and integrating out a and β as follows:

$$\exp\{l_M(y; \tau, \sigma^2)\} \propto \tau^{-1/2} \int \exp\left\{l(y; \beta, a, \sigma^2) - \frac{1}{2\tau} a^T a\right\} da\, d\beta \tag{17}$$

where $l(y; \beta, a, \sigma^2)$ is the log-likelihood (normal) of f under model (15). Wahba (1985) called the maximum marginal likelihood estimator of τ the generalized maximum likelihood estimator. Speed (1991) and Thompson (1985) pointed out that this marginal likelihood (17) of τ is in fact the REML under the linear mixed model

$$y = \mathbf{1}\beta_0 + X\beta_1 + Ba + \epsilon,$$

where $a \sim N(0, \tau I)$ and $\epsilon \sim N(0, \sigma^2 I)$, and B was defined in Sections 3.3–3.4; τ is regarded as a variance component. Hence the maximum marginal likelihood estimator of τ is an REML estimator. The extensive simulation study of Kohn *et al.* (1991) showed that the maximum marginal likelihood estimator of τ has similar and often better performance compared with the generalized cross-validation (GCV) estimator in estimating the nonparametric function.

Zhang *et al.* (1998) extended these researchers' results to estimate the smoothing parameter λ and the variance component θ jointly by using REML for longitudinal data with normally distributed outcomes and a nonparametric mean function. A representative special case of their model can be written as

$$y = f(X) + Zb + \epsilon, \tag{18}$$

where $f(X)$ denotes the values of the nonparametric function $f(\cdot)$ evaluated at the design points of X ($n \times 1$), $b \sim N\{0, D(\theta)\}$ and $\epsilon \sim N\{0, G(\theta)\}$. If $f(\cdot)$ is estimated by using a natural cubic smoothing spline, using equation (8), Zhang *et al.* (1998) rewrote model (18) as a linear mixed model

$$y = \mathbf{1}\beta_0 + X\beta_1 + Ba + Zb + \epsilon, \tag{19}$$

where $a \sim N(0, \tau I)$ and the distributions of b and ϵ are the same as those in model (18). They hence proposed to treat τ as an extra variance component in addition to θ in model (19) and to estimate τ and θ jointly by using REML. Using the results of Harville (1974), this REML corresponds to the marginal likelihood of (τ, θ) constructed by assuming that f takes the form (8) with $a \sim N(0, \tau I)$ and a flat prior for β, and integrating out a and β as follows:

$$\exp\{l_M(y;\tau,\theta)\} \propto |D|^{-1/2}\tau^{-1/2}\int \exp\left\{l(y;\beta,a,b)-\frac{1}{2}b^TD^{-1}b-\frac{1}{2\tau}a^Ta\right\}db\,da\,d\beta, \quad (20)$$

where $l(y;\beta,a,b)=l(y;f,b)$ is the conditional log-likelihood (normal) of f given the random effects b under model (18). Note that the marginal log-likelihood $l_M(y;\tau,\theta)$ in expression (20) has a closed form expression. The simulation results of Zhang *et al.* (1998) show that REML performs very well in estimating both the nonparametric function $f(\cdot)$ and the variance components θ. A similar REML approach was considered by Brumback and Rice (1998) and Wang (1998).

4.2. The marginal quasi-likelihood

In view of the impressive performance of the joint estimation procedure of τ and θ using REML under the Gaussian nonparametric mixed model (18), we propose to extend the marginal likelihood approach of Wahba (1985), Kohn *et al.* (1991) and Zhang *et al.* (1998) to GAMM (1) and to estimate τ and θ jointly by maximizing a marginal quasi-likelihood. Specifically, the GLMM representation (10) suggests that we may treat τ as extra variance components in addition to θ. Similar to the REML (20), we construct the marginal quasi-likelihood of (τ,θ) under GAMM (1) by assuming that f_j takes the form (8) with $a_j\sim N(0,\tau_jI)$ $(j=1,\ldots,p)$ and a flat prior for β and integrating the a_j and β out as follows:

$$\exp\{l_M(y;\tau,\theta)\} \propto |\Lambda|^{-1/2}\int \exp\left\{l(y;\beta,a,\theta)-\frac{1}{2}a^T\Lambda^{-1}a\right\}da\,d\beta$$

$$\propto |D|^{-1/2}|\Lambda|^{-1/2}\int \exp\left\{\sum_{i=1}^{n}-\frac{1}{2\phi}d_i(y;\mu_i^b)-\frac{1}{2}b^TD^{-1}b-\frac{1}{2}a^T\Lambda^{-1}a\right\}db\,da\,d\beta, \quad (21)$$

where $l(y;\beta,a,\theta)=l(y;\beta_0,f_1,\ldots,f_p,\theta)$ was defined in expression (2).

Under the classical nonparametric regression model (15), the marginal quasi-likelihood (21) is simplified as the marginal likelihood (17). Under the Gaussian nonparametric mixed model (18), it reduces to the REML (20). Unlike the Gaussian situation considered in Section 4.1, an evaluation of the marginal quasi-likelihood (21) for non-Gaussian outcomes is hampered by often intractable numerical integration. An approximation is hence proposed in the next section.

4.3. Approximate marginal quasi-likelihood inference

Since the evaluation of the marginal quasi-likelihood $l_M(y;\tau,\theta)$ in expression (21) often involves high dimensional integration, we approximate $l_M(y;\tau,\theta)$ by using Laplace's method. Specifically, taking a quadratic expansion of the exponent of the integrand of expression (21) about its mode before integration and approximating the deviance statistic $d_i(y;\mu_i^b)$ by the Pearson χ^2-statistic (Breslow and Clayton, 1993), derivations similar to those in Appendix A give the approximate marginal log-quasi-likelihood as

$$l_M(y;\tau,\theta)\approx -\tfrac{1}{2}\log|V|-\tfrac{1}{2}\log|X^TV^{-1}X|-\tfrac{1}{2}(Y-X\hat\beta)^TV^{-1}(Y-X\hat\beta), \quad (22)$$

where $V=B\Lambda B^T+ZDZ^T+W^{-1}$. An examination of equation (22) shows that it corresponds to the REML log-likelihood of the working vector Y under the linear mixed model (12) with both a and b as random effects and τ and θ as variance components. It follows that we can easily estimate τ and θ by iteratively fitting model (12) using REML.

Specifically, differentiating expression (22) with respect to $\vartheta = (\tau, \theta)$, some calculation gives the estimating equations for τ and θ, which are the REML equations under the working linear mixed model (12) (also compare Breslow and Clayton (1993), equation (14)),

$$-\frac{1}{2}\operatorname{tr}\left(P\frac{\partial V}{\partial \vartheta_k}\right)+\frac{1}{2}(Y-X\hat{\beta})V^{-1}\frac{\partial V}{\partial \vartheta_k}V^{-1}(Y-X\hat{\beta})=0,$$

which can be equivalently written in terms of the $\hat{f}_j$ as

$$-\frac{1}{2}\operatorname{tr}(PN_jB_jB_j^{\mathrm{T}}N_j^{\mathrm{T}})+\frac{1}{2}\left(Y-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}N_k\hat{f}_k\right)^{\mathrm{T}}R^{-1}N_jB_jB_j^{\mathrm{T}}N_j^{\mathrm{T}}R^{-1}\left(Y-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}N_k\hat{f}_k\right)=0$$

$$-\frac{1}{2}\operatorname{tr}\left(P\frac{\partial R}{\partial \theta_l}\right)+\frac{1}{2}\left(Y-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}N_k\hat{f}_k\right)^{\mathrm{T}}R^{-1}\frac{\partial R}{\partial \theta_l}R^{-1}\left(Y-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}N_k\hat{f}_k\right)=0,$$

where calculations of $\partial V/\partial \vartheta_k$ and $\partial R/\partial \theta_j$ ignore the dependence of W on ϑ and

$$P=V^{-1}-V^{-1}X(X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}=R^{-1}-R^{-1}(X,B)H^{-1}(X,B)^{\mathrm{T}}R^{-1}$$

is the projection matrix under the linear mixed model (12) and $(\hat{\beta}_0, \hat{f}_1, \ldots, \hat{f}_p)$ is the DPQL estimator. For clustered data, it is often computationally efficient to calculate P by using the second expression.

The Fisher information matrix of the approximate marginal quasi-likelihood estimators $\hat{\vartheta} = (\hat{\tau}, \hat{\theta})$ can be approximated by

$$\mathcal{I}(\vartheta)=\begin{pmatrix}\mathcal{I}_{\tau\tau} & \mathcal{I}_{\tau\theta}\\ \mathcal{I}_{\tau\theta}^{\mathrm{T}} & \mathcal{I}_{\theta\theta}\end{pmatrix}, \tag{23}$$

where the (j, k)th element of $\mathcal{I}(\vartheta)$ is $\mathcal{I}_{\vartheta_j\vartheta_k}=0.5\operatorname{tr}(P\,\partial V/\partial\vartheta_j\,P\partial V/\partial\vartheta_k)$. Note that we mainly use equation (23) to construct an approximate covariance matrix of $\hat{\theta}$ and are not interested in using it to make inference on τ. The simulation results of Zhang *et al.* (1998) under the Gaussian nonparametric mixed model (18) show that the estimated standard errors of $\hat{\theta}$ using equation (23) perform very well and are very close to the empirical values. We shall investigate the performance of equation (23) under GAMMs through simulation in Section 7. For terminological simplicity, we call $\hat{\theta}$ and $\hat{\tau}$ the DPQL estimators.

4.4. Summary of inference in generalized additive mixed models under double penalized quasi-likelihood

Calculations in Sections 3.3 and 4.3 suggest that our inference on all model components in GAMMs, including (f_j, θ, τ), can be easily implemented by fitting the working GLMM (10) using the Breslow and Clayton (1993) and Lee and Nelder (1996) penalized quasi-likelihood approach. Equivalently, we only need to fit the working linear mixed model (12) iteratively to the working vector Y, and to use the BLUP estimators of β_j and a_j to construct approximate natural cubic spline estimators $\hat{f}_j$ and to estimate θ and τ by using REML. Hence the existing statistical software SAS macro `GLIMMIX` (Wolfinger, 1996), which repeatedly calls `PROC MIXED`, can be used to estimate (f_j, θ, λ) in GAMMs. A more computationally efficient SAS macro `GAMM` that accounts for the special feature of the GLMM (10) and the linear mixed model (12) is also available from the authors.

5. Bias-corrected double penalized quasi-likelihood

When the data are sparse (e.g. binary), the normal theory based Laplace approximation may not perform well (Lin and Breslow, 1996; Rodríguez and Goldman, 1995). Our simulation study results in Section 7 show that under these circumstances the DPQL estimators of the nonparametric functions $f_j(\cdot)$ often perform well; however, the DPQL estimators of the variance components θ are subject to considerable bias. Since our approximate inference procedure in the GAMM (1) can proceed by fitting the working GLMM (10) using the penalized quasi-likelihood approach of Breslow and Clayton (1993), we propose to apply the GLMM bias correction procedure of Lin and Breslow (1996) to GAMMs with some modifications to obtain better estimates of the variance components.

Specifically, following Lin and Breslow (1996), consider the GAMM with independent random effects

$$g(\mu_i^b) = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \sum_{k=1}^{c} z_{ik}^{\mathrm{T}} b_k, \tag{24}$$

where the random effects b_k ($q_k \times 1$) are independent and follow $b_k \sim N(0, \theta_k I_{q_k})$, and z_{ik} is a $q_k \times 1$ covariate vector. This random effect structure is common in multilevel (hierarchical) studies, such as a multicentre clinical trial, where centre, physician and patient random effects are specified, and in longitudinal studies where random intercepts are specified.

The corrected DPQL estimator of θ is constructed as

$$\hat{\theta}_{\mathrm{C}} = C^{-1} C_{\mathrm{P}} \hat{\theta}, \tag{25}$$

where $\hat{\theta}$ is the DPQL estimator of θ, the correction matrices C and C_{P} are identical with those given in equation (20) of Lin and Breslow (1996), except that their $\hat{\mu}_i^0$ used in calculating W_0, W_1 and W_2 is changed to

$$\hat{\mu}_i^0 = g^{-1}\left\{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{f}_j(x_{ij})\right\},$$

and the $\hat{f}_j(x_{ij})$ are the DPQL estimators. Note that this modified $\hat{\mu}_i^0$ is the estimate of the mean of y_i in equation (24) under independence ($\theta = 0$). The covariance matrix $\hat{\theta}_{\mathrm{C}}$ can be approximated by

$$\mathrm{cov}(\hat{\theta}_{\mathrm{C}}) = (C^{-1} C_{\mathrm{P}})\, \mathrm{cov}(\hat{\theta})(C^{-1} C_{\mathrm{P}})^{\mathrm{T}}.$$

To improve the performance of the DPQL estimators $\hat{f}_j$, we can use $\hat{\theta}_{\mathrm{C}}$ to re-estimate f_j by solving the DPQL estimating equations (11). We evaluate the performance of this correction procedure through simulation in Section 7.

6. Application to infectious disease data

Zeger and Karim (1991) reported longitudinal data on respiratory infection in Indonesian children. 275 preschool childen were examined every quarter up to six consecutive quarters for respiratory infection (0 ≡ no; 1 ≡ yes). The data consisted of 1200 binary observations. The covariates of interest included age in years, xerophthalmia status (0 ≡ no; 1 ≡ yes), an ocular manifestation of chronic vitamin A deficiency, sex (0 ≡ male; 1 ≡ female), height for age and the presence of stunting (0 ≡ no; 1 ≡ yes). For a detailed description of the covariates, see Zeger and Karim (1991).

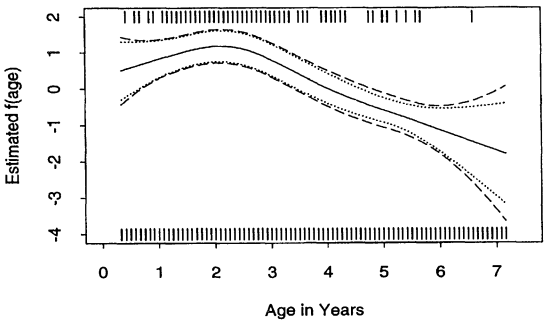

Fig. 1. Estimated DPQL estimate $\hat{f}$(age) (———) for the infectious disease data and its 95% pointwise frequentist (·····) and Bayesian (– – –) confidence intervals: the vertical strokes at 2 and −4 indicate the occurrence of 1s and 0s in the response

Zeger and Karim (1991) considered a parametric logistic–normal model and assumed a linear age effect. An examination of the distribution of the vertical strokes in Fig. 1 suggests that the age effect departs dramatically from linearity. We hence modelled the age effect nonparametrically by using a cubic smoothing spline. Specifically, conditionally on subject-specific random intercepts $b_i \sim N(0, \theta)$, the binary outcomes y_{ij} were assumed to be independent and to follow a semiparametric logistic model

$$\text{logit}\{\text{pr}(y_{ij} = 1|b_i)\} = x_{ij}^{\text{T}}\beta + f(\text{age}_{ij}) + b_i, \tag{26}$$

where j identifies the jth observation of the ith child, x_{ij} contains an intercept, xerophthalmia, seasonal sine and cosine, sex, height and stunted, and $f(\text{age}_{ij})$ is a centred twice-differentiable smooth function of age.

We fitted model (26) using DPQL and corrected DPQL. Note that the semiparametric logistic mixed model (26) differs slightly from GAMM (1) in the extra parametric part $x_{ij}^{\text{T}}\beta$, which can be easily incorporated in the fixed effects part in the working GLMM (10). Therefore, the inference procedure discussed in Sections 3–5 can be used to fit model (26) with trivial modifications.

Since the data contained 83 distinct age values, the amount of computation using equation (13) was minimal. Fig. 1 shows the estimated nonparametric function of age under DPQL and its frequentist and Bayesian 95% confidence intervals. The risk of respiratory infection increased during the first 2 years of life and decreased thereafter. The Bayesian confidence intervals were slightly wider than their frequentist counterparts. The estimated regression coefficients, smoothing parameter and variance component are given in Table 1. The data provided strong evidence for the association between respiratory infection and sex and season. The coefficient of xerophthalmia provided no evidence for vitamin A deficiency on respiratory infection. The Bayesian standard errors of the regression coefficients were slightly larger than the frequentist standard errors. The bias-corrected variance component estimate was larger than the uncorrected quantity. The simulation study in Section 7 shows that the bias-corrected variance component estimate is often associated with smaller bias.

7. Simulation study

We conducted a simulation study to evaluate the performance of DPQL and corrected DPQL. Each data set was composed of 100 clusters of size $n_i = 5$. Conditionally on the

Table 1. Parameter estimates for the infectious disease data

Parameter	*Estimate*	*Bayes standard error*	*Frequentist standard error*
Intercept	−2.92	0.24	0.23
Vitamin A	0.52	0.46	0.46
Seasonal cosine	−0.58	0.17	0.17
Seasonal sine	−0.16	0.17	0.17
Sex	−0.50	0.24	0.24
Height for age	−0.03	0.02	0.02
Stunted	0.39	0.43	0.42
τ	0.27	0.32	
θ (DPQL)	0.38	0.26	
θ (corrected DPQL)	0.48	0.33	

cluster-specific random intercepts $b_i \sim N(0,\ \theta = 0.5)$, the binary observations y_{ij} were generated within each cluster with conditional probabilities ($i = 1, \ldots, 100, j = 1, \ldots, 5$)

$$\text{logit}\{E(y_{ij}|b_i)\} = \beta_0 + \beta_1 t_i + f_1(x_{2i}) + f_2(x_{3ij}) + b_i,$$

where $\beta_0 = -0.5$, $\beta_1 = 1$, t_i takes value 1 for half of the subjects and 0 for the other and mimics a binary treatment indicator, and

$$f_1(x_1) = \frac{1}{3}\{2F_{8,8}(x_1) + F_{5,5}(x_1)\} - 1,$$

$$f_2(x_2) = \frac{1}{10}\{6F_{30,17}(x_2) + 4F_{3,11}(x_2)\} - 1,$$

$$F_{p,q}(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1},$$

and $\Gamma(\cdot)$ is a gamma function. Here $f_1(x_1)$ is a unimodal function, whereas $f_2(x_2)$ is a bimodal function. Similar test functions were used in Wahba (1983).

We assumed x_1 to be a cluster level covariate with 50 equally spaced knots in [0, 1] and x_2 to be a covariate varying within each cluster with 100 equally spaced knots in [0, 1]. Specifically, we let $x_{1i} = \text{trun}\{(i+1)/2\}/100$ and

$$x_{2ij} = \frac{\text{trun}\{(i+4)/5\}}{100} + 0.20(j-1)$$

for $i = 1, \ldots, 100$ and $j = 1, \ldots, 5$, where trun$(\cdot)$ denotes a truncation operator. The constants 1 used in $f_1(x_1)$ and $f_2(x_2)$ were chosen to centre f_1 and f_2. In other words, the means of f_1 and f_2 over the distinct values of x_1 and x_2 were 0. 500 data sets with 500 observations each were generated and analysed using DPQL and corrected DPQL. The entire experiment was repeated with the binomial observations y_{ij} whose denominator was 8.

Table 2 gives the estimated fixed effects $\hat{\beta}_0$ and $\hat{\beta}_1$, and their empirical and estimated frequentist and Bayesian standard errors (SEs) under DPQL for binomial denominators $m = 1$ and $m = 8$. For both scenarios, DPQL worked well in estimating the fixed effects. The estimated fixed effects had little biases. Both frequentist and Bayesian SEs agreed well with the empirical SEs. The Bayesian SEs were larger than their frequentist counterparts. The frequentist SEs were slightly closer to the empirical SEs.

Table 2. Means and standard errors of fixed effects estimates over 500 replications†

m	*Parameter*	*Mean*	*Empirical SE*	*Estimated SE, frequentist*	*Estimated SE, Bayesian*
1	β_0	−0.52	0.18	0.18	0.18
	β_1	0.98	0.25	0.26	0.28
8	β_0	−0.54	0.11	0.11	0.12
	β_1	1.06	0.16	0.17	0.19

†True values: $\beta_0 = -0.5$; $\beta_1 = 1.0$.

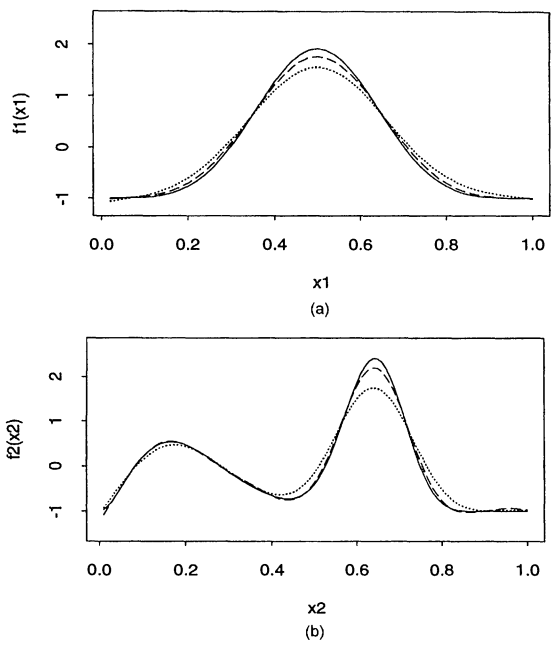

Fig. 2. True and estimated nonparametric functions (a) $\hat{f}_1(x_1)$ and (b) $\hat{f}_2(x_2)$ based on 500 replications: ———, true; ·····, $m = 1$; — — —, $m = 8$

Fig. 2 depicts the true curves $f_j(x_j)$ $(j = 1, 2)$ and the estimated curves $\hat{f}_j(x_j)$ under DPQL for the binomial denominator equal to 1 and 8 respectively. DPQL overall performed well in estimating the nonparametric functions. The estimated nonparametric functions had noticeable negative biases when the data were binary ($m = 1$) and approached the true values quickly as the binomial denominator became 8. The biases were more pronounced at the values of x where the curvatures were high. This suggests that DPQL slightly oversmooths the curves for binary data. This might be due to the fact that the Laplace approximation (22) to the marginal quasi-likelihood might not perform well for sparse binary data and the resultant smoothing parameters τ were underestimated.

Fig. 3 compares the empirical and estimated pointwise Bayesian and frequentist SEs. Both SEs agree well with the empirical SEs. The Bayesian SEs were slightly larger than the frequentist counterparts. The frequentist SEs were closer to the empirical SEs when the bias of the estimated nonparametric function was small. The Bayesian SEs agreed better with the empirical SEs when the bias became more visible.

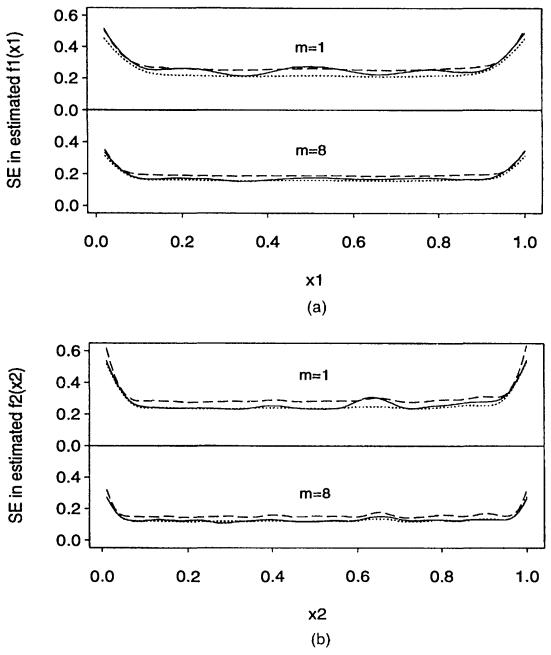

Fig. 3. Empirical, frequentist and Bayesian pointwise standard errors of estimated nonparametric functions (a) $\hat{f}_1(x_1)$ and (b) $\hat{f}_2(x_2)$ based on 500 replications: ———, empirical SE; ·····, frequentist SE; — — —, Bayesian SE

Fig. 4 compares the empirical pointwise coverage probabilities of the 95% confidence intervals of $f_1(x_1)$ and $f_2(x_2)$ calculated using the frequentist and Bayesian SEs. For binary data ($m = 1$), the coverage probabilities of both the frequentist and the Bayesian confidence intervals were very close to 95% at the values of x where the bias was small. At these places of x, the coverage probabilities of the frequentist confidence intervals agreed slightly better with the nominal value, whereas the coverage probabilities of the Bayesian confidence intervals were slightly larger than the nominal value. When the bias became visible, both confidence intervals yielded unsatisfactory coverage probabilities and the Bayesian confidence intervals performed slightly better. This is probably because the Bayesian SEs account for the bias in the estimated nonparametric function. A plausible explanation of slight overshooting of the coverage probabilities of the Bayesian confidence intervals when the bias is small is that the Bayesian SEs account for the global bias in the estimated nonparametric function. When the global bias is visible, the Bayesian confidence intervals may overcompensate for the bias at the values of x where the bias is small. As the binomial denominator m increased to 8, the coverage probabilities quickly approached the nominal value.

The overall performances of the frequentist and Bayesian confidence intervals were comparable and the Bayesian intervals had slightly superior performance. For $m = 1$, the means of the coverage probabilities of the frequentist and Bayesian confidence intervals over the range of x_1 were 84.0% and 89.4% for $f_1(x_1)$, and their means over the range of x_2 were 85.7% and 90.8% for $f_2(x_2)$. When $m = 8$, both confidence intervals performed well and their coverage probabilities quickly approached the nominal value. The means of the coverage probabilities of the frequentist and Bayesian confidence intervals were 90.9% and 95.3% for $f_1(x_1)$ and 91.3% and 96.3% for $f_2(x_2)$.

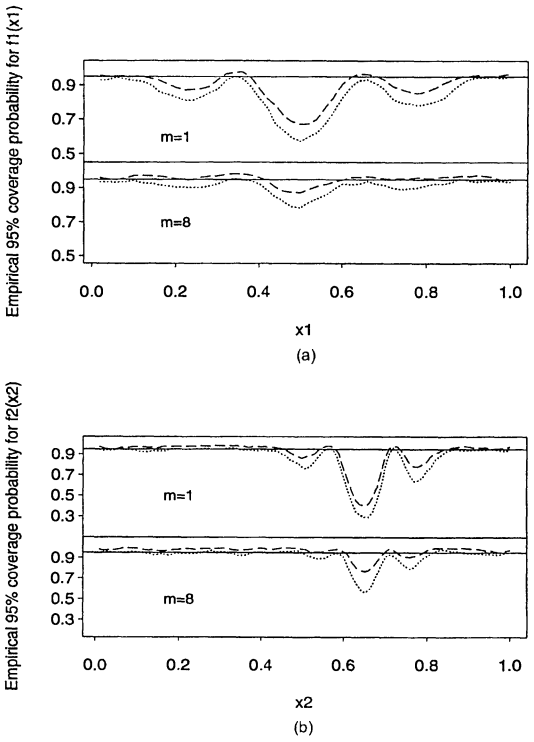

Fig. 4. Estimated pointwise frequentist and Bayesian 95% coverage probabilities of the values of the true fixed functions (a) $f_1(x_1)$ and (b) $f_2(x_2)$ based on 500 replications: ———, nominal level 95%; ·····, coverage probability of the frequentist confidence interval; — — —, coverage probability of the Bayesian confidence interval

Table 3. Means and SEs of variance component estimates over 500 replications†

m	*Method*	*Mean*	*Empirical SE*	*Estimated SE*	*Mean-square error*
1	DPQL	0.33	0.18	0.21	0.06
	Corrected DPQL	0.42	0.22	0.26	0.05
8	DPQL	0.46	0.09	0.09	0.01
	Corrected DPQL	0.47	0.09	0.09	0.01

†True value: $\theta = 0.5$.

Table 3 gives the estimated variance components, their empirical and estimated SEs and mean-square errors under DPQL and corrected DPQL for binomial denominators $m = 1$ and $m = 8$. The DPQL method considerably underestimated the variance component for binary data ($m = 1$). The bias correction factor significantly reduced the bias. Although the variance was inflated, the bias-corrected estimate had a smaller mean-square error. A similar performance of the bias correction factor was found with GLMMs (Lin and Breslow, 1996). As the sample size increases, we expect that the bias would become more dominated and the reduction in mean-square error would be more significant. When m increased to 8, the performance of both the uncorrected, and the corrected variance component estimates quickly improved and the bias was negligible. Although we may obtain better estimates of the nonparametric functions by using the corrected variance component estimate to re-

estimate f_1 and f_2 by solving equation (13), the improvement was minimal and the results are not presented.

8. Discussion

We have proposed in this paper nonparametric regression for correlated data within the framework of GAMMs. Smoothing splines were used to estimate the nonparametric functions and marginal quasi-likelihood was used to make simultaneous inference on the smoothing parameters and the variance components. Because of the often required high dimensional integration, DPQL was proposed to make approximate inference. A key feature of this approach is that all model components of GAMMs, including the nonparametric functions, the smoothing parameters and the variance components, can be estimated in a unified parametric mixed model framework. Specifically, we only need to fit a working GLMM by repeatedly fitting a linear mixed model to a modified outcome variable.

Our simulation studies show that DPQL performs well in estimating the nonparametric functions. The frequentist and Bayesian confidence intervals are comparable. For correlated binary data, which is a worst case scenario for the approximate inference procedure, both confidence intervals provide approximately correct coverage probabilities, except for the places where the curvatures of the nonparametric functions are high and the biases in the estimated nonparametric functions are visible. As the binomial denominator increases, the coverage probabilities of both confidence intervals quickly approach the nominal value. Although the variance component estimates under DPQL are subject to considerable bias when the data are binary, the simple bias correction procedure significantly reduces the bias and yields practically satisfactory estimates when the variance components are small or moderate (Lin and Breslow, 1996). As expected, the accuracy quickly improves as the binomial denominator increases.

We have discussed in this paper approximate inference using DPQL and corrected DPQL. Similarly to the performance of the penalized quasi-likelihood approach in parametric GLMMs (Breslow and Clayton, 1993; Lin and Breslow, 1996), except for the Gaussian case, the estimates under DPQL are often asymptotically biased in GAMMs, especially for binary data. Although the correction procedure reduces the bias, we expect that it will not work well when the variance components are large and it is not applicable to GAMMs with correlated random effects. It is hence of interest to compute full cubic smoothing spline estimators of the nonparametric functions in GAMMs by directly maximizing the penalized quasi-likelihood (3) and full marginal quasi-likelihood estimators of the smoothing parameters and the variance components by directly maximizing expression (21). Since they both require evaluating high dimensional integration, especially expression (21), one may use Monte Carlo simulation methods, such as Gibbs sampling (Zeger and Karim, 1991) and Metropolis sampling (McCulloch, 1997). These methods are particularly attractive when the variance components are large and the data are sparse. However, the dimension of integration required by these procedures may be overwhelming. The practical feasibility of these Monte Carlo simulation methods hence deserves careful consideration.

We have discussed in this paper the use of the marginal quasi-likelihood to estimate the smoothing parameters. Alternative approaches to selecting the smoothing parameters include cross-validation and GCV (Rice and Silverman, 1991; Wahba, 1985). A challenge in using GCV is that it is so far not defined under GAMMs. Compared with cross-validation, our marginal quasi-likelihood inference procedure has several advantages:

(a) it requires much less computation;
(b) inference on the variance components is incorporated naturally;
(c) it is applicable to clustered and crossed designs.

Further research is needed to compare the performance of the estimated nonparametric functions using cross-validation and GCV (to be defined) with that using the marginal quasi-likelihood.

In this paper, we have focused on spline smoothing. It is of interest to investigate in future research kernel smoothing for estimating the nonparametric functions in GAMMs. Unlike spline smoothing, which allows us to estimate all model components in a unified parametric mixed model framework, a challenge in kernel smoothing in GAMMs is that new methods need to be developed for the estimation of the bandwidth parameters and the variance components.

Acknowledgements

This work was supported in part by a grant from the US National Cancer Institute and grants from the University of Michigan.

Appendix A: Derivation of expression (5)

Equation (2) takes the form $\int \exp\{-\kappa(b)\}\,db$. Making a quadratic expansion of $-\kappa(b)$ about its maximum point $\tilde{b}$ before integration and plugging the resultant approximation into equation (3), penalized quasi-likelihood (3) can be approximated by

$$-\frac{1}{2}\log|I+Z^{\mathrm{T}}\tilde{W}ZD|-\frac{1}{2\phi}\sum_{i=1}^{n}d_i(y_i;\mu_i^{\tilde{b}})-\frac{1}{2}\tilde{b}^{\mathrm{T}}D^{-1}\tilde{b}-\frac{1}{2}\sum_{j=1}^{p}\lambda_j f_j^{\mathrm{T}}K_j f_j, \tag{27}$$

where $\tilde{b}=\tilde{b}(\beta_0, f_1, \ldots, f_p, \theta)$ satisfies

$$Z^{\mathrm{T}}W\Delta(y-\mu^b)-D^{-1}b=0,$$

and $\Delta=\mathrm{diag}\{g'(\mu_i^b)\}$, $W=\mathrm{diag}[\{\phi m_i^{-1}v(\mu_i^b)g'(\mu_i^b)^2\}^{-1}]$ and $\tilde{W}=W(\mu^{\tilde{b}})$.

Assuming that $\tilde{W}$ varies slowly with $(\beta_0, f_1, \ldots, f_p)$ for fixed θ (Breslow and Clayton, 1993), we may ignore the first term in expression (27) when maximizing it with respect to $(\beta_0, f_1, \ldots, f_p)$. Simple calculations show that the resultant estimators $(\hat{\beta}_0, \hat{f}_1, \ldots, \hat{f}_p)$ can be equivalently obtained by maximizing the DPQL in expression (5) jointly with respect to $(\beta_0, f_1, \ldots, f_p)$ and b.

References

Aitkin, M. (1998) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, to be published.

Berhane, K. and Tibshirani, R. J. (1996) Generalized additive models for longitudinal data. *Can. J. Statist.*, to be published.

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

Brumback, B. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Ass.*, **93**, 961–1006.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*, 2nd edn. New York: Wiley.

Green, P. J. (1987) Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.*, **55**, 245–260.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Härdle, W. (1990) *Applied Non-parametric Regression*. Cambridge: Cambridge University Press.

Hart, J. D. (1991) Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B, **53**, 173–187.

Harville, D. A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Kohn, R., Ansley, C. F. and Tharm, D. (1991) The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Am. Statist. Ass.*, **86**, 1042–1050.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.
Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Lin, X. and Breslow, N. E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Statist. Ass.*, **91**, 1007–1016.
McCulloch, C. (1997) Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–190.
O'Sullivan, F., Yandell, B. S. and Raynor, W. J. (1986) Automatic smoothing of regression functions in generalized linear models. *J. Am. Statist. Ass.*, **81**, 96–103.
Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.
Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *J. R. Statist. Soc.* A, **158**, 73–89.
Speed, T. (1991) Discussion on BLUP is a good thing: the estimation of random effects, by Robinson, G. K. *Statist. Sci.*, **6**, 15–51.
Thompson, R. (1985) Discussion on Some aspects of the spline smoothing approach to non-parametric regression curve fitting (by B. W. Silverman). *J. R. Statist. Soc.* B, **47**, 43–44.
Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.
Verbyla, A. P. (1995) A mixed model formulation of smoothing splines and test linearity in generalized linear model. *Technical Report 95/5*. Department of Statistics, University of Adelaide, Adelaide.
Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1998) The analysis of designed experiments and longitudinal data using smoothing splines. Submitted to *Appl. Statist.*
Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B, **40**, 364–372.
———(1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.
———(1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, **13**, 1378–1402.
Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc.* B, **60**, 159–174.
Wild, C. J. and Yee, T. W. (1996) Additive extensions to generalized estimating equation methods. *J. R. Statist. Soc.* B, **58**, 711–725.
Wolfinger, R. (1996) The GLIMMIX SAS macro. Cary: SAS Institute.
Zeger, S. L. and Diggle, P. J. (1994) Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.
Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Statist. Ass.*, **86**, 79–86.
Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998) Semi-parametric stochastic mixed models for longitudinal data. *J. Am. Statist. Ass.*, **93**, 710–719.