

# Likelihood and Conditional Likelihood Inference for Generalized Additive Mixed Models for Clustered Data

Daowen Zhang\* and Marie Davidian

Department of Statistics, North Carolina State University

Box 8203, Raleigh, North Carolina 27695-8203, U.S.A.

\**Email:* dzhang2@stat.ncsu.edu *Tel:* (919) 515-1933 *Fax:* (919) 515-7591

## ABSTRACT

Lin and Zhang [1] proposed the generalized additive mixed model (GAMM) as a framework for analysis of correlated data, where normally distributed random effects are used to account for correlation in the data, and proposed to use double penalized quasi-likelihood (DPQL) to estimate the nonparametric functions in the model and marginal likelihood to estimate the smoothing parameters and variance components simultaneously. However, the normal distributional assumption for the random effects may not be realistic in many applications, and it is unclear how violation of this assumption affects ensuing inferences for GAMMs. For a particular class of GAMMs, we propose a conditional estimation procedure built on a conditional likelihood for the response given a sufficient statistic for the random effect, treating the random effect as a nuisance parameter, which thus should be robust to its distribution. In extensive simulation studies, we assess performance of this estimator under a range of conditions and use it as a basis for comparison to DPQL to evaluate the impact of violation of the normality assumption. The procedure is illustrated with application to data from the Multicenter AIDS Cohort Study (MACS).

**KEY WORDS:** Longitudinal data; Marginal likelihood; Nonparametric regression; Smoothing spline; Variance component.

# 1 Introduction

Clustered data arise frequently in biomedical research endeavors such as epidemiology and clinical trials. For example, each subject in a longitudinal epidemiological study or each hospital in a multi-center clinical trial may be viewed as a cluster. The challenge in analyzing clustered data is that the data within a cluster tend to be correlated. A popular way to account for this feature is to use cluster-specific random effects to model the correlation explicitly in a generalized linear mixed model (GLMM). Under a normal or other parametric distributional assumption for the random effects, likelihood inference can be carried out using a Monte Carlo approach or numerical integration (Zeger and Karim [2], Booth and Hobert [3]). When the random effects structure is complex, full likelihood inference may not be feasible. In this case, approximate inference using penalized quasi-likelihood approach of Breslow and Clayton [4] or a fully Bayesian approach [5] is usually adopted.

In many situations, however, the usual parametric linear assumption for the fixed covariate effects in a GLMM may not be an appropriate representation of the true underlying relationship between covariates and the response of interest. Lin and Zhang [1] considered an extension of GLMMs, generalized additive mixed models (GAMMs), where additive non-parametric functions are used to model this relationship. They formulated the nonparametric functions using smoothing splines and estimated the nonparametric functions by maximizing a double penalized quasi-likelihood (DPQL). Using the mixed model representation of a GAMM, they cast the estimation and inference in a GLMM framework and estimated the smoothing parameters jointly with variance components of the random effects by treating the inverses of the smoothing parameters as extra variance components.

As for most popular mixed models, the Lin and Zhang [1] approach to inference in GAMMs is based on the potentially strong assumption that the random effects are normally distributed. Because the random effects in a GAMM represent variation of cluster-specific

characteristics, the normal distribution may be too restrictive to represent the true features of this underlying variation. Thus, it is of considerable interest to understand the effect of violation of this assumption on performance of DPQL for these models and to develop alternative procedures for GAMMs that do not require normality of the random effects.

In mixed models with parametric covariate effects, two main approaches to achieving this latter goal have been proposed. The first is to relax the assumption on the random effects and represent their distribution directly in a likelihood framework. Magder and Zeger [6] proposed a smooth nonparametric maximum likelihood approach; Tao et al. [7] estimated the density of a scalar random effect via their predictive recursive algorithm; Verbeke and Lesaffre [8] used a mixture of normals to model the random effects, which they implement via an EM algorithm (Verbeke and Molenberghs [9]); and Zhang and Davidian [10] considered a semi-nonparametric (SNP) density representation for the random effects. Chen, Zhang and Davidian [11] extended the SNP methodology to generalized linear mixed model and implemented maximum likelihood inference via Monte Carlo EM algorithm. Aitkin [12] considered EM-based nonparametric maximum likelihood approach for GLMMs. The second approach is to treat the random effects as nuisance parameters and base inference on a conditional likelihood; e.g., Verbeke, Spiessens and Lesaffre [13] considered conditional inference for linear mixed models, while Jiang [14] developed conditional inference for GLMMs where the likelihood is based on a subset of the random effects.

In this paper, we propose a conditional inference procedure for GAMMs by treating the random effects as nuisance parameters and estimating the nonparametric functions by maximizing a penalized conditional likelihood. Similar to the DPQL approach of Lin and Zhang [1], the smoothing parameters are estimated using a mixed effect representation of a nonparametric function. We describe the model specification in Section 2. In Section 3 we develop estimation procedures for the nonparametric functions in the model and the

smoothing parameters. In Section 4, we report on extensive simulation studies to investigate the performance of this estimator under a range of conditions and use it as a basis for comparison to DPQL to assess the effect of departures from normality of the random effects. We illustrate the new procedure with application to data from the Multicenter AIDS Cohort Study (MACS) in Section 5.

## 2 Model Specification

Consider a random sample of  $m$  independent clusters, where, for cluster  $i$ ,  $i = 1, \dots, m$ , we observe responses  $y_{ij}$ ,  $j = 1, \dots, n_i$ , and values for  $p$  corresponding covariates  $x_{1ij}, \dots, x_{pij}$  that vary within cluster  $i$ . Conditional on a cluster-specific random effect  $b_i$  and the covariates, the  $y_{ij}$  are assumed to be independent and have exponential family density

$$f(y_{ij}|b_i) = e^{\{y_{ij}\eta_{ij}^b - h(\eta_{ij}^b)\}/a_{ij}(\phi) + c(y_{ij}, \phi)}, \quad (1)$$

where  $h(\cdot)$  and  $c(\cdot, \cdot)$  are known functions,  $a_{ij}(\phi) = \phi/\omega_{ij}$  and  $\omega_{ij}$  is a known prior weight (such as the denominator of a binomial distribution), and  $\phi$  is a dispersion parameter. As the dispersion parameter is known and equal to unity in several popular cases, such as the binomial and Poisson, we focus on the situation  $\phi = 1$  in the sequel.

Let  $\mu_{ij}^b = E(y_{ij}|b_i)$ , the conditional mean of  $y_{ij}$  given  $b_i$ . In this paper, we consider the particular GAMM with random intercept only for  $\mu_{ij}^b$ , given by

$$g(\mu_{ij}^b) = \eta_{ij}^b = f_1(x_{1ij}) + \dots + f_p(x_{pij}) + b_i, \quad (2)$$

where  $g(\cdot)$  is the canonical link function,  $f_v(\cdot)$ ,  $v = 1, \dots, p$ , are centered, twice-differential smooth but arbitrary functions, and the intercept is absorbed into the random intercept  $b_i$ . Different from Lin and Zhang [1], who assumed the random effect  $b_i$  is normally distributed, we do not impose any distributional assumption on  $b_i$ . Interest focuses on making inference on the functions  $f_v(\cdot)$ .

For each  $v = 1, \dots, p$ , let  $X_v^0$  be the  $r_v$ -dimensional vector of distinct values (knots) of the  $x_{vij}$ , and let  $f_v$  be the unknown vector of the values of  $f_v(\cdot)$  evaluated at  $X_v^0$ . Denote by  $N_v$  the incidence matrix mapping the  $x_{vij}$  to  $X_v^0$  and  $N_{vi}$  the  $i$ th block corresponding to the  $i$ th cluster, and write  $\mu_i^b$  for the conditional mean vector of  $y_i = (y_{i1}, \dots, y_{in_i})^T$  given  $b_i$ . Then model (2) can be re-written in matrix notation as

$$g(\mu_i^b) = \eta_i^b = N_{1i}f_1 + \dots + N_{pi}f_p + 1_{n_i}b_i, \quad (3)$$

where  $1_{n_i}$  is a  $n_i$ -vector of ones. Write  $\eta_i^b = (\eta_{i1}^b, \dots, \eta_{in_i}^b)^T$ .

### 3 Estimation Procedure

#### 3.1 Estimation of nonparametric functions

We develop a conditional estimation procedure for the nonparametric functions in (2) by treating the cluster-specific random effects  $b_i$  as nuisance parameters. From the form of the conditional distribution of  $y_{ij}$  given  $b_i$  in (1) and the assumption on  $\mu_{ij}^b$  in (2), it is easy to show that  $s_i = \sum_{j=1}^{n_i} y_{ij} = y_{i+}$  is a sufficient and complete statistic for  $b_i$  and that the conditional distribution of  $y_i$  given  $s_i$  is

$$f(y_i|s_i) = e^{y_i^T \omega_i (N_{1i}f_1 + \dots + N_{pi}f_p) - G_i(f_1, \dots, f_p; y_i)},$$

where  $\omega_i$  is a  $n_i \times n_i$  diagonal matrix with  $j$ th diagonal element  $\omega_{ij}$  (or  $\omega_{ij}/\phi$  if  $\phi \neq 1$ ), and  $G_i(f_1, \dots, f_p; y_i)$  is a function of  $f_1, \dots, f_p$  and  $y_i$  only. Calculation of  $G_i(f_1, \dots, f_p; y_i)$  is straightforward for given  $f_1, \dots, f_p$  and  $y_i$ . For example, if the response  $y_{ij}$  is discrete, then

$$G_i(f_1, \dots, f_p; y_i) = \log \left\{ \sum_{u_{i+}=s_i} e^{u_i^T \omega_i (N_{1i}f_1 + \dots + N_{pi}f_p) + \sum_{j=1}^{n_i} c(u_{ij}, \phi)} \right\},$$

where the summation is over all possible  $u_i = (u_{i1}, \dots, u_{in_i})^T$  in the sample space such that  $u_{i+} = \sum_{j=1}^{n_i} u_{ij} = s_i$ .

Therefore, a conditional log-likelihood for  $(f_1, \dots, f_p)$  for given data  $y = (y_1^T, \dots, y_m^T)^T$  is

$$\ell_c(f_1, \dots, f_p; y) = \sum_{i=1}^m \log\{f(y_i|s_i)\} = \sum_{i=1}^m y_i^T \omega_i (N_{1i}f_1 + \dots + N_{pi}f_p) - \sum_{i=1}^m G_i(f_1, \dots, f_p; y_i).$$

Because each  $f_v(\cdot)$  is an infinite-dimensional parameter, we propose to estimate nonparametric functions  $f_v(\cdot)$  by maximizing the penalized conditional likelihood

$$\ell_{pc}\{f_1(\cdot), \dots, f_p(\cdot), \lambda_1, \dots, \lambda_p; y\} = \ell_c(f_1, \dots, f_p; y) - \sum_{v=1}^p \frac{\lambda_v}{2} \int \{f_v''(x)\}^2 dx,$$

where the integral  $\int \{f_v''(x)\}^2 dx$  measures the roughness of the nonparametric function  $f_v(\cdot)$ , and the  $\lambda_v$  are positive smoothing parameters controlling goodness-of-fit of the model to the data and roughness of the  $f_v(\cdot)$ . In the special case where the  $f_v(\cdot)$  are linear functions and  $y_{ij}$ 's are clustered binary responses, we obtain the conditional logistic regression model.

Because  $\ell_c(f_1, \dots, f_p; y)$  depends on the unknown functions  $f_v(\cdot)$  only through the  $f_v$ , the values of the  $f_v(\cdot)$  evaluated at the corresponding distinct knots, it follows immediately from Green and Silverman [15] or Zhang et al. [16] that the estimates of the  $f_v(\cdot)$  are natural cubic smoothing splines, and there exist  $p$  semi-positive definite matrices  $K_v$  of rank  $q_v = r_v - 2$  such that  $\int \{f_v''(x)\}^2 dx = f_v^T K_v f_v$ . For each  $v = 1, \dots, p$ , decompose  $K_v$  as  $K_v = L_v L_v^T$ , where  $L_v$  is a  $r_v \times q_v$  full rank matrix. Suppose each covariate is centered such that  $1_{r_v}^T X_v^0 = 0$ . Then  $f_v$  can be expressed as  $f_v = X_v^0 \beta_v + B_v a_v$ , where  $B_v = L_v (L_v^T L_v)^{-1}$ . Under this parameterization,  $f_v$  automatically satisfies  $1_{r_v}^T f_v = 0$  (i.e.,  $f_v(\cdot)$  is centered). Let  $X_i = (N_{1i}X_1^0, \dots, N_{pi}X_p^0)$ ,  $\beta = (\beta_1, \dots, \beta_p)^T$ ,  $Z_i = (N_{1i}B_1, \dots, N_{pi}B_p)$  and  $a = (a_1^T, \dots, a_p^T)^T$ . Then the conditional likelihood can be re-parameterized as

$$\ell_c(f_1, \dots, f_p; y) = \ell_c(\beta, a; y) = \sum_{i=1}^m \tilde{y}_i^T (X_i \beta + Z_i a) - \sum_{i=1}^m G_i(\beta, a; y_i),$$

where  $\tilde{y}_i = \omega_i y_i$ , and the penalized conditional likelihood becomes

$$\ell_{pc}(\beta, a, \lambda; y) = \ell_c(\beta, a; y) - \sum_{v=1}^p \frac{\lambda_v}{2} a_v^T a_v, \quad (4)$$

where  $\lambda = (\lambda_1, \dots, \lambda_p)^T$  is the vector of smoothing parameters.

For given  $\lambda$ , taking derivatives of  $\ell_{pc}(\beta, a, \lambda; y)$  with respect to  $\beta$  and  $a$  and setting them equal to zero, we obtain estimating equations for  $\beta$  and  $a$  given by

$$\begin{cases} X^T(\tilde{y} - \tilde{\mu}^c) = 0 \\ Z^T(\tilde{y} - \tilde{\mu}^c) - \Lambda a = 0, \end{cases} \quad (5)$$

where  $X$  and  $Z$  are the matrices obtained by stacking  $X_i$  and  $Z_i$  for  $i = 1, \dots, m$ , respectively;  $\tilde{\mu}^c = E(\tilde{y}|s)$ ;  $\tilde{y} = (\tilde{y}_1^T, \dots, \tilde{y}_m^T)^T$ ;  $s = (s_1, \dots, s_m)^T$ ; and  $\Lambda$  is a block-diagonal matrix with  $v$ th block equal to  $\lambda_v I_{q_v \times q_v}$  and  $I_{q_v \times q_v}$  is the  $q_v \times q_v$  identity matrix.

The estimating equations (5) can be solved iteratively using Newton-Raphson algorithm. Given current estimates  $\beta^{(0)}, a^{(0)}$ , expand  $\tilde{\mu}^c$  as

$$\tilde{\mu}^c \approx \tilde{\mu}_0^c + W[X, Z] \begin{bmatrix} \beta - \beta^{(0)} \\ a - a^{(0)} \end{bmatrix},$$

where  $\tilde{\mu}_0^c$  is  $\tilde{\mu}^c$  evaluated at  $(\beta^{(0)}, a^{(0)})$ ,  $W = \text{diag}\{W_i\}$ , and  $W_i = \text{var}(\tilde{y}_i|s_i)$ . Then the equations (5) become

$$\begin{cases} X^T \tilde{y} = X^T \tilde{\mu}_0^c + X^T W X (\beta - \beta^{(0)}) + X^T W Z (a - a^{(0)}) \\ Z^T \tilde{y} = Z^T \tilde{\mu}_0^c + Z^T W X (\beta - \beta^{(0)}) + Z^T W Z (a - a^{(0)}) + \Lambda a, \end{cases}$$

which leads to the Newton-Raphson update

$$\begin{bmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z + \Lambda \end{bmatrix} \begin{bmatrix} \beta^{(1)} \\ a^{(1)} \end{bmatrix} = \begin{bmatrix} X^T Y \\ Z^T Y \end{bmatrix}, \quad (6)$$

where  $Y = \tilde{y} - \tilde{\mu}_0^c + W X \beta^{(0)} + W Z a^{(0)}$ , and the algorithm is iterated until convergence. Denote by  $\hat{\beta}$  and  $\hat{a}$  the solution at convergence. Then  $f_v$  is estimated by  $\hat{f}_v = X_v^0 \hat{\beta}_v + B_v \hat{a}_v$ , which can be used to determine the entire function  $f_v(\cdot)$ .

Note that in the case of Gaussian response, equation (6) can be solved without iteration. In this case, the  $i$ th block of  $Y$  and  $W$  are given by  $Y_i = (y_i - 1_{n_i} y_{i+}/n_i)/\phi$  and  $W_i =$

$(I_{n_i \times n_i} - \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T / n_i) / \phi$ . Iteration is necessary for other cases and computation can become potentially intensive. However, the items needed in (6) can be calculated easily, at least conceptually. For example, in the case of binary response, we basically only need to calculate

$$\tilde{\mu}_i^c = E(\tilde{y}_i | s_i) = \frac{\sum_{u_{i+}=s_i} u_i e^{u_i^T \eta_i}}{\sum_{u_{i+}=s_i} e^{u_i^T \eta_i}}, \quad W_i = \text{var}(\tilde{y}_i | s_i) = \frac{\sum_{u_{i+}=s_i} u_i u_i^T e^{u_i^T \eta_i}}{\sum_{u_{i+}=s_i} e^{u_i^T \eta_i}} - \tilde{\mu}_i^c (\tilde{\mu}_i^c)^T$$

where  $\eta_i = N_{1i} f_1 + \dots + N_{pi} f_p$ , and the summation is over all possible  $u_i$ , a  $n_i$ -vector of 0's and 1's, such that  $u_{i+} = s_i$ . Obviously, the computation could be intensive if  $n_i$  is large.

Because in this scheme the estimation of the nonparametric functions in the model is built on a penalized conditional likelihood given a sufficient and complete statistic for the random effect, the estimated nonparametric functions do not depend on the random effect nor on its distribution. Therefore, if we have good estimates of the smoothing parameters, intuitively, the nonparametric function estimates should have good statistical properties such as robustness to the random effect distribution, small bias and accurate coverage properties.

Denote the coefficient matrix on the left hand side of the system (6) at convergence by  $H$ , which is the same as the negative derivative of the score vector in (5) with respect to  $\beta$  and  $a$ . Then the variance of  $\hat{\beta}$  and  $\hat{a}$  can be approximated by

$$\text{var} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = H^{-1} \text{var} \begin{bmatrix} X^T(\tilde{y} - \tilde{\mu}^c) \\ Z^T(\tilde{y} - \tilde{\mu}^c - \Delta a) \end{bmatrix} \Big|_{(\tilde{y}|s)} H^{-1} = H^{-1} \begin{bmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z \end{bmatrix} H^{-1}. \quad (7)$$

Then, as  $\hat{f}_v = X_v^0 \hat{\beta}_v + B_v \hat{a}_v$ , an approximation to the variance of  $\hat{f}_v$  can be derived easily from this expression, and the  $(1-\alpha)$ th point-wise confidence intervals of  $f_v(\cdot)$  can be approximately constructed as  $\hat{f}_v \pm z_{\alpha/2} \text{SE}(\hat{f}_v)$ , where  $\text{SE}(\hat{f}_v)$  is the estimated standard errors of  $\hat{f}_v$ , which is calculated as the squared-root of the diagonal elements of the estimated variance matrix of  $\hat{f}_v$ .

Note that the Newton-Raphson update (6) can be viewed as the mixed model equations

for the mixed model for the working vector  $Y$  given by

$$Y = WX\beta + WZa + \epsilon, \quad (8)$$

where  $\beta$  is the fixed effect,  $a \sim N(0, \Lambda^{-1})$  is the random effect and  $\epsilon \sim N(0, W)$ . This is similar to the mixed model representation of a GAMM discussed in Lin and Zhang [1].

### 3.2 Estimation of smoothing parameters

Lin and Zhang [1] considered estimation of smoothing parameters for the nonparametric functions in a GAMM using a marginal likelihood approach based on the mixed model representation of a GAMM. Because we have a similar mixed model representation (8), we propose to estimate the smoothing parameters  $\lambda_v$ ,  $v = 1, \dots, p$ , in a similar way by treating  $a_v \sim N(0, \lambda_v^{-1}I_{q_v \times q_v})$ . To take into account estimation of  $\beta$ , we consider estimating the  $\lambda_v$  by maximizing the marginal conditional likelihood

$$L_M(\lambda; y) = |\Lambda|^{1/2} \int e^{\ell_{pc}(\beta, a, \lambda; y)} d\beta da.$$

The integration in the above likelihood does not have a closed form expression except in the case of Gaussian response and is often numerically intractable because the high integration dimension (i.e.,  $\sum_{v=1}^p r_v - 1$ ) prohibits any attempt for direct evaluation. Following Lin and Zhang [1], we evaluate  $L_M(\lambda; y)$  by Laplace approximation. For a given  $\lambda$ , denote explicitly by  $\hat{\beta}(\lambda)$  and  $\hat{a}(\lambda)$  the mode of  $\ell_{pc}(\beta, a, \lambda; y)$ , i.e., the solution of  $\beta$  and  $a$  given in equation (6) in Section 3.1 at convergence. Then  $L_M(\lambda; y)$  can be approximated by

$$L_M(\lambda; y) \approx |\Lambda|^{1/2} |H|^{-1/2} e^{\ell_{pc}\{\hat{\beta}(\lambda), \hat{a}(\lambda), \lambda; y\}},$$

where  $H$  is the coefficient matrix of the system (6) at convergence. We thus suggest estimating  $\lambda$  by maximizing the approximate log marginal conditional likelihood function

$$\ell_M(\lambda; y) \approx \frac{1}{2} \sum_{v=1}^p q_v \log(\lambda_v) - \frac{1}{2} \log |H| + \ell_{pc}\{\hat{\beta}(\lambda), \hat{a}(\lambda), \lambda; y\}. \quad (9)$$

The approximate log marginal conditional likelihood function (9) of  $\lambda$  can be maximized using the Newton-Raphson algorithm. To use this algorithm, it is necessary to calculate the first and second derivatives of  $\ell_M(\lambda; y)$  with respect to  $\lambda$ . As  $\{\hat{\beta}(\lambda), \hat{a}(\lambda)\}$  maximize  $\ell_{pc}(\beta, a, \lambda; y)$  for any given  $\lambda$ , we have for  $v = 1, \dots, p$  by the chain rule

$$\begin{aligned} \frac{\partial \ell_{pc}\{\hat{\beta}(\lambda), \hat{a}(\lambda), \lambda; y\}}{\partial \lambda_v} &= \frac{\partial \ell_{pc}(\beta, a, \lambda; y)}{\partial(\beta^T, a^T)} \Big|_{\{\hat{\beta}(\lambda), \hat{a}(\lambda)\}} \begin{bmatrix} \frac{\partial \hat{\beta}(\lambda)}{\partial \lambda_v} \\ \frac{\partial \hat{a}(\lambda)}{\partial \lambda_v} \end{bmatrix} + \frac{\partial \ell_{pc}(\beta, a, \lambda; y)}{\partial \lambda_v} \Big|_{\{\hat{\beta}(\lambda), \hat{a}(\lambda)\}} \\ &= -\frac{1}{2} \hat{a}_v^T(\lambda) \hat{a}_v(\lambda). \end{aligned}$$

Thus, the first derivative of  $\ell_M(\lambda; y)$  is given by

$$\frac{\partial \ell_M(\lambda; y)}{\partial \lambda_v} = \frac{q_v}{2\lambda_v} - \frac{1}{2} \text{tr} \left( H^{-1} \frac{\partial H}{\partial \lambda_v} \right) - \frac{1}{2} \hat{a}_v^T(\lambda) \hat{a}_v(\lambda).$$

Denote the block of  $H^{-1}$  corresponding to  $(\lambda_v, \lambda_{v'})$  by  $H^{v,v'}$  for  $v, v' = 1, \dots, p$ . If we assume that the conditional variance  $W$  varies with  $\lambda$  slowly so that we can ignore the dependence of  $W$  on  $\lambda$ , then the above derivative can be simplified to

$$\frac{\partial \ell_M(\lambda; y)}{\partial \lambda_v} = \frac{q_v}{2\lambda_v} - \frac{1}{2} \text{tr}(H^{vv}) - \frac{1}{2} \hat{a}_v^T(\lambda) \hat{a}_v(\lambda). \quad (10)$$

Taking derivatives of this expression with respect to  $\lambda_v$  and  $\lambda_{v'}$  and ignoring the dependence of  $W$  on  $\lambda$  again leads to the second derivatives

$$\begin{aligned} \frac{\partial^2 \ell_M(\lambda; y)}{\partial \lambda_v^2} &= -\frac{q_v}{2\lambda_v^2} + \frac{1}{2} \text{tr}\{(H^{vv})^2\} - \hat{a}_v^T(\lambda) \frac{\partial \hat{a}_v(\lambda)}{\partial \lambda_v} \\ \frac{\partial^2 \ell_M(\lambda; y)}{\partial \lambda_v \partial \lambda_{v'}} &= \frac{1}{2} \text{tr}(H^{vv'} H^{v'v}) - \hat{a}_v^T(\lambda) \frac{\partial \hat{a}_v(\lambda)}{\partial \lambda_{v'}}, \end{aligned}$$

for  $v, v' = 1, \dots, p$ . Using the derivative rule for an implicit function, it is easily shown that

$$\frac{\partial \hat{a}_v(\lambda)}{\partial \lambda_{v'}} = -H^{vv'} \hat{a}_{v'}(\lambda),$$

for any  $v, v' = 1, \dots, p$ . Hence the second derivatives needed for Newton-Raphson algorithm are given by

$$\frac{\partial^2 \ell_M(\lambda; y)}{\partial \lambda_v^2} = -\frac{q_v}{2\lambda_v^2} + \frac{1}{2} \text{tr}\{(H^{vv})^2\} + \hat{a}_v^T(\lambda) H^{vv} \hat{a}_v(\lambda) \quad (11)$$

$$\frac{\partial^2 \ell_M(\lambda; y)}{\partial \lambda_v \partial \lambda_{v'}} = \frac{1}{2} \text{tr}(H^{vv'} H^{v'v}) + \hat{a}_v^T(\lambda) H^{vv'} \hat{a}_{v'}(\lambda). \quad (12)$$

Denote by  $S(\lambda)$  the first derivative given in (10) and by  $I(\lambda)$  the  $p \times p$  matrix with the  $(v, v')$ th element being the negative of the second derivatives in (11) and (12). Then the Newton-Raphson algorithm proceeds until convergence by iteration of the update

$$\lambda^{(1)} = \lambda^{(0)} + I^{-1}(\lambda^{(0)}) S(\lambda^{(0)}), \quad (13)$$

where  $\lambda^{(0)}$  is an initial estimate for  $\lambda$ . Our experience shows that iterating between (13) and (6) works well and may be computationally more efficient.

For numerical stability, we may use Fisher-scoring type of algorithm to maximize  $\ell_M(\lambda; y)$  under the mixed model representation (8) by treating  $a_v \sim N(0, \lambda_v^{-1} I_{q_v \times q_v})$ . Denote the block of  $H^{-1}$  corresponding to  $a$  by  $H^{aa}$ . Then it is easy to show that  $\text{var}(\hat{a}) = \Lambda^{-1} - H^{aa}$  under this distributional assumption for  $a$ , and we have simple expectations for the second derivatives given in (11) and (12)

$$\mathbb{E} \left( \frac{\partial^2 \ell_M(\lambda; y)}{\partial \lambda_v^2} \right) = -\frac{1}{2} \text{tr}(\lambda_v^{-1} I - H^{vv})^2, \quad \mathbb{E} \left( \frac{\partial^2 \ell_M(\lambda; y)}{\partial \lambda_v \partial \lambda_{v'}} \right) = -\frac{1}{2} \text{tr}(H^{vv'} H^{v'v}).$$

Then replacing the elements of  $I(\lambda)$  by their expectations given above yields the Fisher scoring algorithm. Again, we may iterate between (13) and (6) until convergence.

## 4 Simulation Study

We conducted extensive simulation studies to evaluate the performance of the DPQL estimation procedure of Lin and Zhang [1] and that of the conditional estimation procedure under different distributional assumptions and different magnitudes of the variance of the random effect in the model. The conditional method moreover serves as a benchmark for performance that can be achieved if the normality assumption is relaxed, so that compari-

son to DPQL highlights how possible nonrobustness of DPQL to nonnormality may manifest itself and the extent to which improvement is possible.

For each cluster  $i = 1, \dots, 500$ , conditionally independent binary responses  $y_{ij} \sim \text{Bin}(1, \pi_{ij}^b)$  ( $j = 1, \dots, 5$ ) were generated from the GAMM

$$\text{logit}(\pi_{ij}^b) = f(x_{ij}) + b_i, \quad (14)$$

where  $x_{ij} = \text{trun}\{(i + 24)/25\}/100 + 0.2(j - 1)$  (i.e., every group of 25 clusters has the same set of covariate values of  $x$ ), and  $f(x)$  is defined by

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

for  $F_{p,q}(x)$  a Beta density function with parameters  $p$  and  $q$ . Five distributions of  $b_i$  were considered: (1) Normal,  $b_i \sim N(0, 0.5)$ ; (2) Mixture of normals,  $b_i \sim 0.7N(-0.42, 0.0884) + 0.3N(0.98, 0.0884)$ ; (3)  $t$ -distribution with 5 degrees of freedom; (4)  $\chi^2$  distribution with 1 degree of freedom; and (5) Bernoulli distribution with success probability 0.2. The  $b_i$  were linearly transformed so that they all have mean zero and the same variance 0.5. One hundred data sets were generated for each case and the DPQL of Lin and Zhang [1] and the conditional estimation procedure developed in Section 3 were applied to each data set. The simulation was repeated for those five distributions by scaling the random effects so that their variances equal to 1 and 2. For numerical stability, the covariate  $x$  was multiplied by 20.

Because the proposed conditional estimation procedure conditions on the sum of the responses, those clusters with responses all equal to 0 or 1 are automatically removed from the analysis. When the variance of the random effect is 0.5, about 50 clusters (10%) were removed. The numbers of such clusters went up to about 75 (15%) for variance 1 and about 100 (20%) for variance 2. The conditional estimation algorithm did not reach convergence for about 5 simulated data sets in every 100 simulation runs. The DPQL of Lin and Zhang

[1] converged for all data sets. The comparison was based on the data sets where both estimation procedures converged.

Table 1 presents the average of the estimated smoothing parameters for the different simulation scenarios for both methods and shows that those from the conditional estimation procedure are very stable and are consistently smaller than the DPQL estimates. Because a smaller smoothing parameter corresponds to a less-smooth nonparametric function estimate and the DPQL estimate of the nonparametric function tends to over-smooth the underlying function, the results in this table imply that the proposed conditional estimation procedure may produce a less biased estimate of the underlying function.

Table 2 presents the average of 100 (biased-corrected) estimates of the variance of the random effect  $b_i$ , denoted by  $\theta$ , using the DPQL estimation procedure, which treats  $b_i$  as normally distributed, as well as the Monte Carlo standard deviation of the estimated sampling variances and the average of 100 estimated standard errors of the variance estimates. The estimated variances are reasonably close to the true values for small-to-moderate variance components (0.5, 1) for all distributions except for  $t$ -distribution with 5 degrees of freedom and  $\chi^2$  distribution with 1 degree of freedom. When the true value of  $\theta$  increases to 2, it is severely underestimated in all cases. Surprisingly, the estimated standard errors and the Monte Carlo standard deviation of the estimated variances agree with each other very well.

Figures 1–3 present the true and the average of the estimated nonparametric functions, mean squared errors, and empirical coverage probabilities of 95% point-wise confidence intervals using both the DPQL and conditional procedures for different simulation designs. A notable feature of these figures is the robustness of the DPQL estimation procedure to the misspecification of the distribution of the random effect, especially for small-to-moderate variance component  $\theta$ . When  $\theta$  is as large as 2, although the estimated nonparametric functions are still able to capture the overall shape of the underlying true function, the estimates

are over-smoothed, and hence cannot estimate well the peaks in the true function. This results in large mean squared errors and low coverage probabilities near the peaks. In contrast, the proposed conditional estimation procedure performs consistently better than the DPQL procedure in that it yields less-biased nonparametric function estimates, similar or smaller mean squared errors and better coverage properties, especially near the peaks. One reason for this observation may be that the bias in estimates of the random effect variances induces bias in the nonparametric function estimates. The other reason may be that DPQL involves two types of approximation, one for the random effect and one for the smoothing parameters, while conditional approach only involves one approximation for the smoothing parameter. Therefore, the conditional approach eliminates any biases introduced by the approximation for the random effect when using DPQL. However, as in the binary response case studied here a potential drawback of the proposed approach is the need to eliminate some data from the analysis, and because inference is built on a conditional likelihood, the estimated nonparametric functions are more variable in some regions, and the difference becomes little larger with the increase of the variance of the random effect since more data are removed in this case using the conditional approach.

## 5 An Example

In this section, we illustrate the proposed estimation procedure through application to data from the MACS study [17]. The human immune deficiency virus (HIV) weakens or destroys the immune system by attacking CD4+ cells, which perform critical functions in coordinating the body's immune response. Accordingly, the number of CD4+ cells (CD4 count) is used routinely to monitor disease progression in HIV-infected individuals. CD4 counts range from 500 to 1500+ cells/mm<sup>3</sup> blood, and, typically, a CD4 count below 500 cells/mm<sup>3</sup> is taken as

evidence of impaired immunologic status that may place the patient at risk of opportunistic infection. In MACS, a total of 2376 CD4 count measurements were collected from a cohort of 369 men infected with HIV, and one objective was to examine how the probability of experiencing CD4 count below 500 cells/mm<sup>3</sup> changes over the course of seroconversion; i.e., the period during which a patient is discerned to have developed detectable antibodies as the result of HIV infection.

Denote by  $y_{ij}$  for subject  $i$  at the  $j$  time point the binary variable indicating whether or not the CD4 count of subject  $i$  is below 500 (1=yes, 0=no), and by  $t_{ij}$  the years since seroconversion. Given subject-specific random effect  $b_i$ , let  $\mu_{ij}^b = P[y_{ij} = 1|b_i]$ . To characterize the probability of CD4 < 500 over time, we consider the following special generalized additive mixed model

$$\text{logit}(\mu_{ij}^b) = f(t_{ij}) + b_i, \quad (15)$$

where  $f(t)$  is a centered smooth nonparametric function of time since seroconversion. No parametric distribution is assumed for  $b_i$  when using our new approach. However a normal distribution  $N(\alpha, \theta)$  is assumed when the DPQL of Lin and Zhang [1] is used.

Figures 4(a) and 4(b) present the estimated nonparametric function  $\hat{f}(t)$  and the corresponding 95% confidence intervals using the DPQL and the conditional estimation procedure. Overall, they look similar and hence yield the similar qualitative conclusion that the probability of having CD4 below 500 is quite stable a few months before seroconversion. This probability increases sharply until about one year after seroconversion and then increases more gradually. However, the estimated nonparametric functions differ quantitatively between the methods, especially after seroconversion: The estimated function using DPQL has smaller rates of change than that from the conditional estimation procedure. This difference is probably due to the large variance component  $\theta$ , which is estimated to be  $\hat{\theta} = 2.46$ . From the simulation studies presented in Section 4, the nonparametric function estimate using

DPQL may exhibit larger biases when the random effect has a variance of this magnitude. The smoothing parameter was estimated to be 1.3 for DPQL and 0.9 for the conditional estimation procedure. The mean of the random effect  $b_i$  was estimated to be  $\hat{\alpha} = -1.24$  with estimate standard error 0.12 using DPQL.

## 6 Discussion

In this paper, we have proposed a conditional estimation procedure for a GAMM for clustered data with the canonical link as an alternative to the DPQL estimation procedure of Lin and Zhang [1]. The conditional estimation procedure is built on the conditional distribution of the response given a sufficient and complete statistic for the random effect, treating it as a nuisance parameter, and is hence robust to any random effect distribution. Our simulation results indicate, interestingly, that the DPQL estimation procedure, which assumes normal random effect, is very robust to this assumption for estimating the nonparametric function and the variance component of the random effect as long as the true variance of the random effect is not too large. For large random effect variance, the estimated nonparametric function may suffer from large bias, especially near peaks of the true function, and the estimated variance component may be severely under-estimated. In contrast, the estimated nonparametric function from the proposed conditional estimation procedure has consistently smaller bias and better coverage properties, demonstrating that improvements over DPQL are possible if the normality assumption can be relaxed.

Although the proposed conditional estimation procedure shows favorable performance over the DPQL estimation procedure, it has some inherent disadvantages. First, the proposed method only applies to a GAMM with the canonical link, which may hinder its application. Second, due to the nature of the method, some data have to be removed from the

analysis, and the final inference is conditioned on the sum of the response. Hence the gain of robustness and smaller biases comes with the loss of efficiency, and the effects of any cluster-level covariates cannot be estimated. Third, although the proposed estimation procedure eliminates the need for numerical integration, it can also be computationally expensive for certain types of response (i.e., binomial data with large binomial denominator or Poisson data with large  $n_i$ ). Forth, it is less stable compared to DPQL and may fail to converge for some data sets. Nonetheless, viewing the method as a basis for comparison to illuminate possible shortcomings of DPQL suggests that future research on procedures for GAMMs that do not require parametric assumptions on the random effects would be valuable.

## ACKNOWLEDGMENTS

The first author's research was supported by the NIH grant R01 CA085848 and the second author's research was supported in part by NIH grants R01 CA085848 and R37 AI031789.

## REFERENCES

- [1] X. Lin, D. Zhang, Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society, Series B*, **61** (1999) 381-400.
- [2] S.L. Zeger, M.R. Karim, Generalized linear models with random effects: A Gibbs sampling approach, *Journal of American Statistical Association*, **86** (1991) 79-86.
- [3] J.G. Booth, J.P. Hobert, Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society, Series B*, **61** (1999) 265-285.
- [4] N.E. Breslow, D.G. Clayton, Approximate inference in generalized linear mixed models, *Journal of American Statistical Association*, **88** (1993) 9-25.

- [5] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, London: Chapman and Hall, 1995.
- [6] L.S. Magder, S.L. Zeger, A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians, *Journal of American Statistical Association*, **91** (1996) 1141-1151.
- [7] H. Tao, M. Palta, B.S. Yandell, M.A. Newton, An estimation method for the semiparametric mixed effects model, *Biometrics*, **55** (1999) 102-110.
- [8] G. Verbeke, E. Lesaffre, A linear mixed-effects model with heterogeneity in the random-effects population, *Journal of the American Statistical Association*, **91** (1996), 217-221.
- [9] G. Verbeke, G. Molenberghs, *Linear Mixed Models for Longitudinal Data*, New York: Springer, 2000.
- [10] D. Zhang, M. Davidian, Linear mixed models with flexible distributions of random effects for longitudinal data, *Biometrics*, **57** (2001) 795-802.
- [11] J. Chen, D. Zhang, M. Davidian, A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution, *Biostatistics*, **3** (2002) 347-360.
- [12] M. Aitkin, A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics*, **55** (1999) 117-128.
- [13] G. Verbeke, B. Spiessens, E. Lesaffre, Conditional Linear Mixed Models, *American Statistician*, **55** (2001) 25-34.
- [14] J.M. Jiang, Conditional inference about generalized linear mixed models, *Annals of Statistics*, **27** (1999) 1974-2007.

- [15] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Models, London: Chapman and Hall, 1994.
- [16] D. Zhang, X. Lin, J. Raz, M. Sowers, Semiparametric stochastic mixed models for longitudinal data, Journal of the American Statistical Association, **93** (1998) 710-719.
- [17] R.A. Kaslow, D.G. Ostrow, R. Detels *et al.*, The Multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants, American Journal of Epidemiology, **126** (1987), 310-318.

**Table 1**

*Comparison of estimated  $\lambda$ 's using Lin and Zhang's [1] DPQL and the proposed conditional marginal likelihood (CML) approach;  $\theta$  is the true variance of the random effect*

Distribution	$\theta = 0.5$		$\theta = 1$		$\theta = 2$	
	DPQL	CML	DPQL	CML	DPQL	CML
Normal	2.3	1.9	2.5	1.9	2.8	1.9
Normal mixture	2.4	2.0	2.4	1.9	2.4	1.8
$t_5$	2.4	2.0	2.6	1.9	2.8	2.0
$\chi_1^2$	2.2	1.9	2.2	1.9	2.3	1.8
Bernoulli	2.3	2.0	2.2	1.9	2.2	1.8

**Table 2**

*Estimated variance of random effect using Lin and Zhang's [1] DPQL; Ave. is the Monte Carlo average of the variance estimates, SD is the standard deviation of the estimates and SE is the Monte Carlo average of the estimated standard errors based on 100 simulation runs;  $\theta$  is the true variance of the random effect*

Distribution	$\theta = 0.5$			$\theta = 1$			$\theta = 2$		
	Ave.	SD	SE	Ave.	SD	SE	Ave.	SD	SE
Normal	0.46	0.08	0.10	0.84	0.12	0.13	1.52	0.18	0.18
Normal mixture	0.46	0.09	0.10	0.90	0.14	0.13	1.75	0.19	0.19
$t_5$	0.40	0.10	0.10	0.74	0.12	0.12	1.27	0.16	0.16
$\chi_1^2$	0.36	0.09	0.10	0.59	0.12	0.11	1.00	0.17	0.14
Bernoulli	0.44	0.11	0.10	0.87	0.14	0.13	1.46	0.19	0.17

## LIST OF FIGURES

**Figure 1.** True and estimated nonparametric functions, mean squared errors and empirical coverage probabilities of 95% point-wise confidence intervals using DPQL of Lin and Zhang [1] and the conditional estimation procedure based on 100 simulation runs for  $\theta = 0.5$ : —, true; - - - -, DPQL; - - - -, conditional procedure. The distributions in 5 rows are (1) Normal; (2) Mixture of normals; (3)  $t_5$ ; (4)  $\chi_1^2$ ; (5) Bernoulli.

**Figure 2.** True and estimated nonparametric functions, mean squared errors and empirical coverage probabilities of 95% point-wise confidence intervals using DPQL of Lin and Zhang [1] and the conditional estimation procedure based on 100 simulation runs for  $\theta = 1$ : —, true; - - - -, DPQL; - - - -, conditional procedure. The distributions in 5 rows are (1) Normal; (2) Mixture of normals; (3)  $t_5$ ; (4)  $\chi_1^2$ ; (5) Bernoulli.

**Figure 3.** True and estimated nonparametric functions, mean squared errors and empirical coverage probabilities of 95% point-wise confidence intervals using DPQL of Lin and Zhang [1] and the conditional estimation procedure based on 100 simulation runs for  $\theta = 2$ : —, true; - - - -, DPQL; - - - -, conditional procedure. The distributions in 5 rows are (1) Normal; (2) Mixtures of normals; (3)  $t_5$ ; (4)  $\chi_1^2$ ; (5) Bernoulli.

**Figure 4.** Estimated nonparametric function and its 95% confidence intervals for  $f(t)$  in model (15) using DPQL of Lin and Zhang [1] **(a)** and the conditional estimation procedure **(b)**.

Figure 1

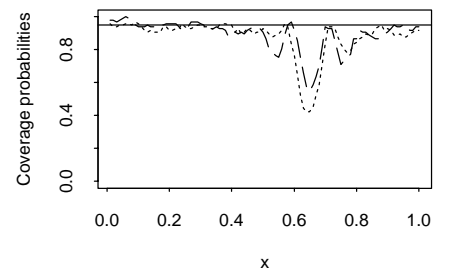
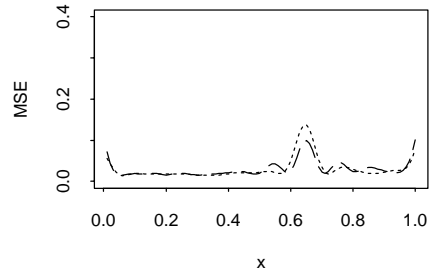
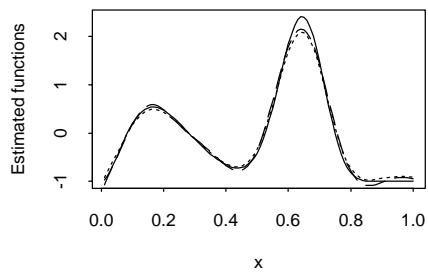
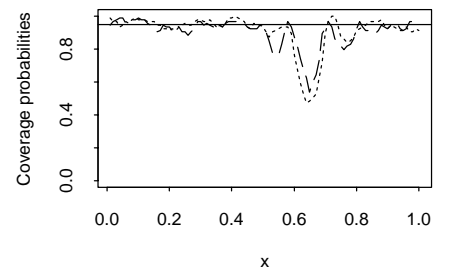
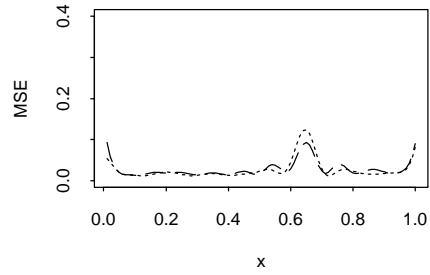
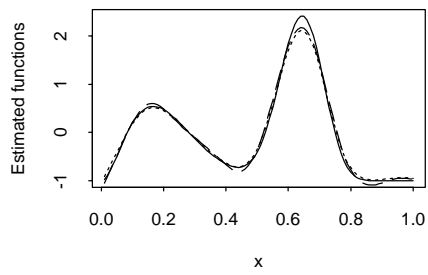
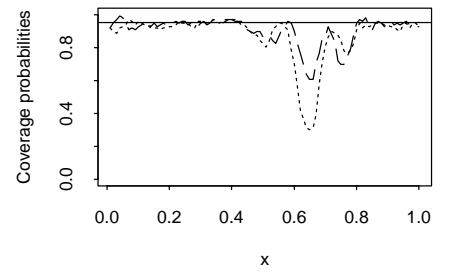
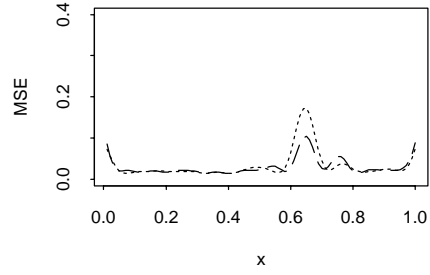
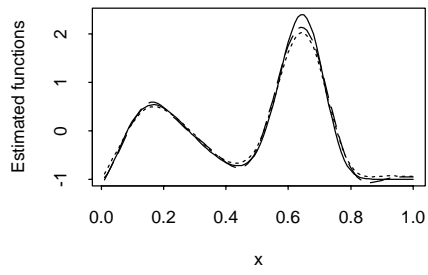
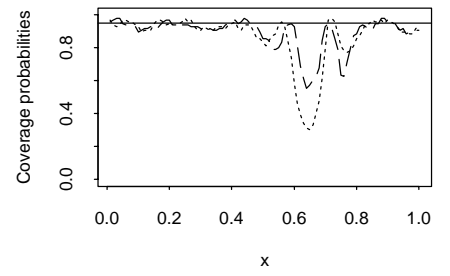
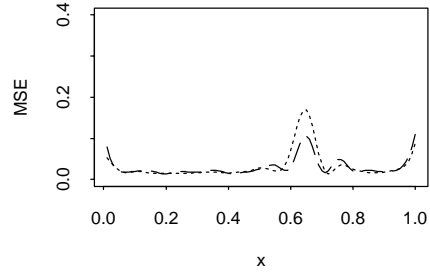
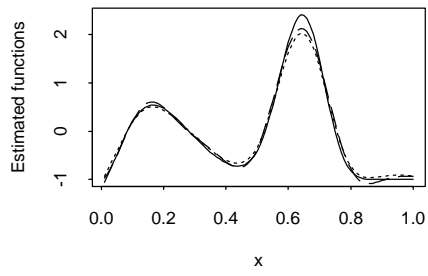
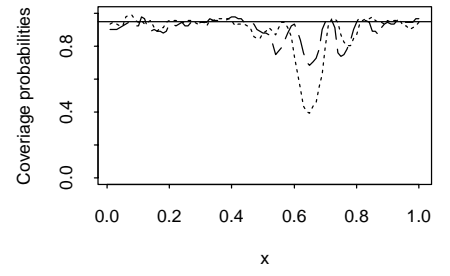
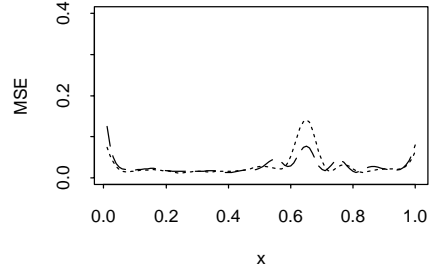
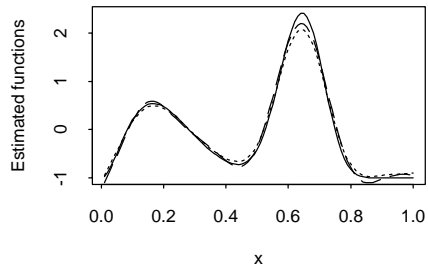


Figure 2

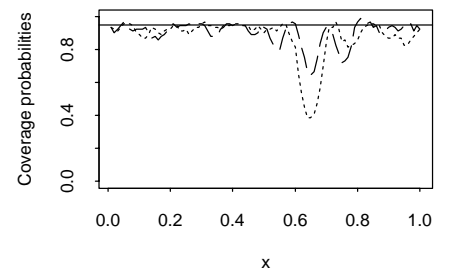
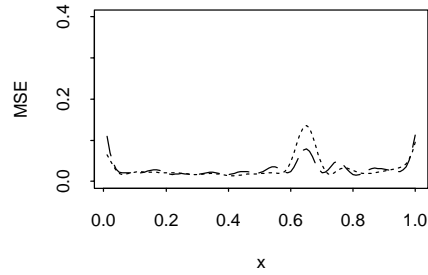
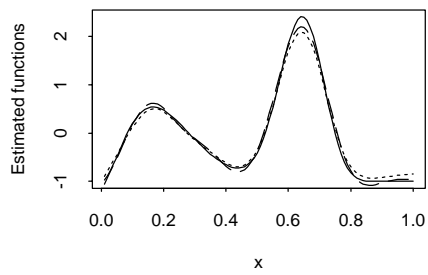
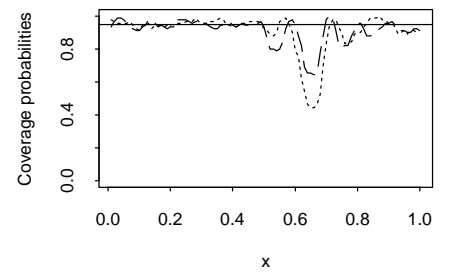
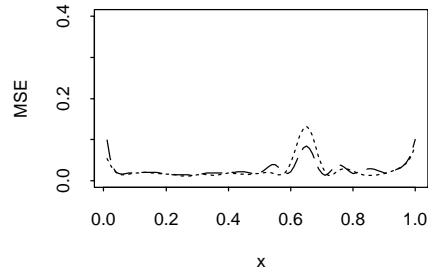
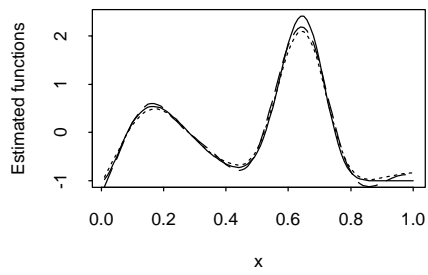
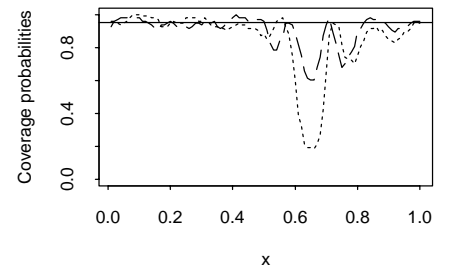
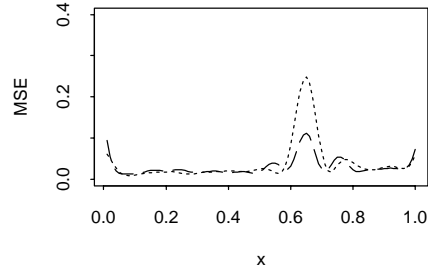
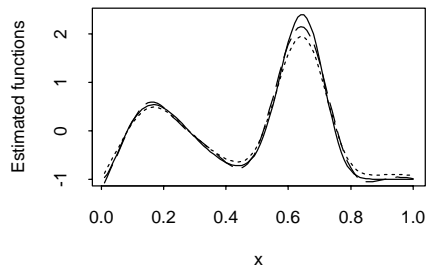
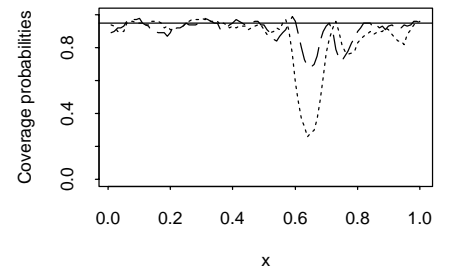
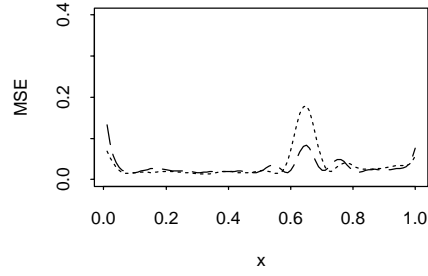
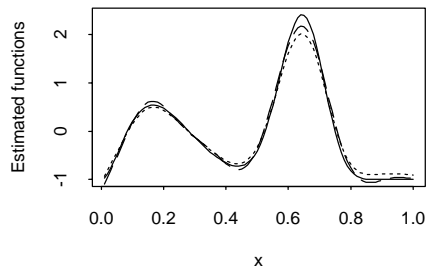
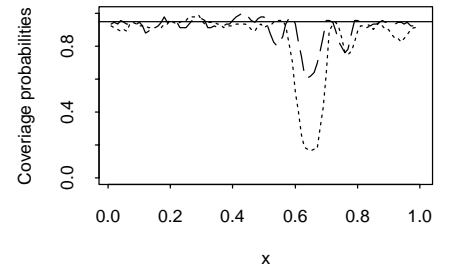
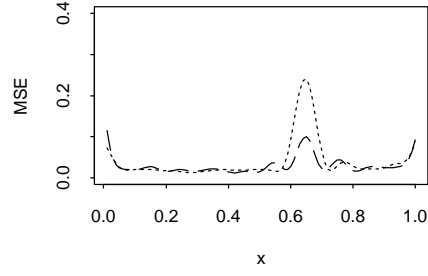
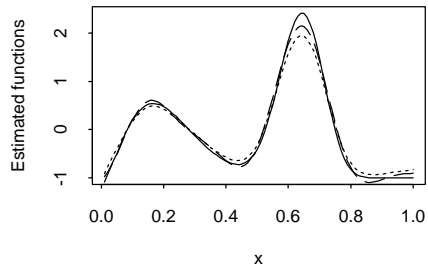


Figure 3

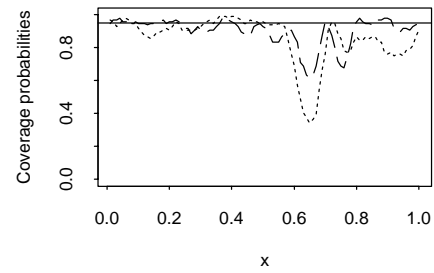
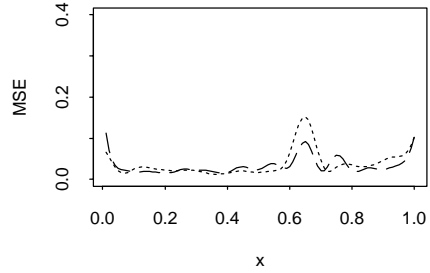
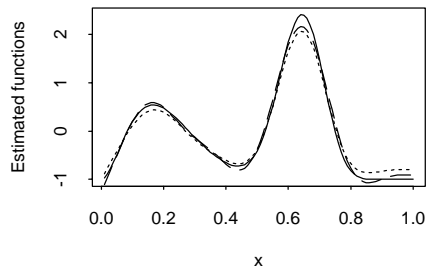
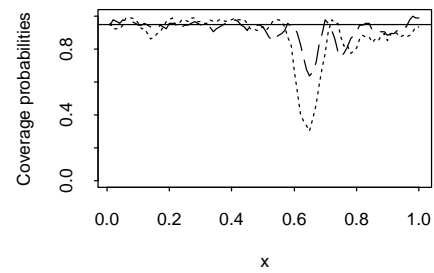
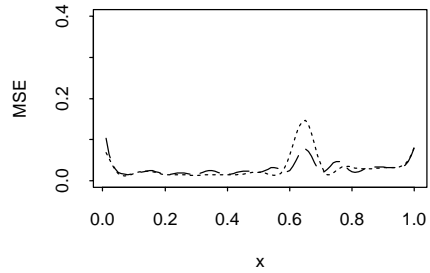
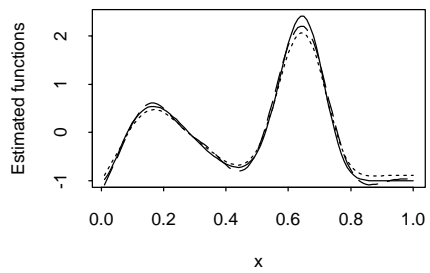
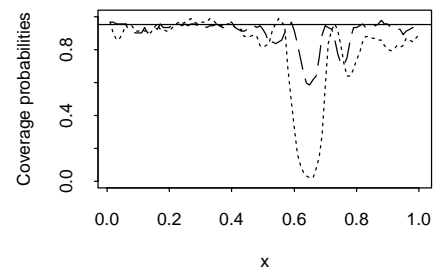
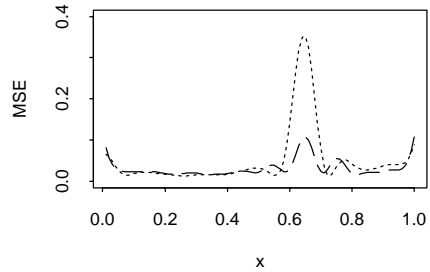
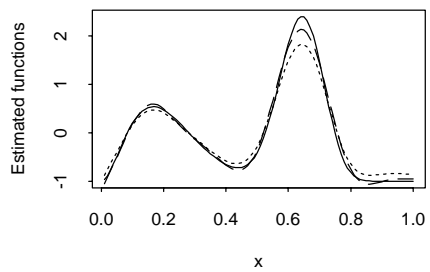
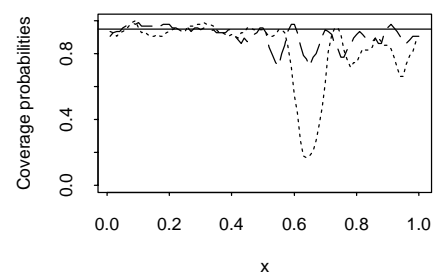
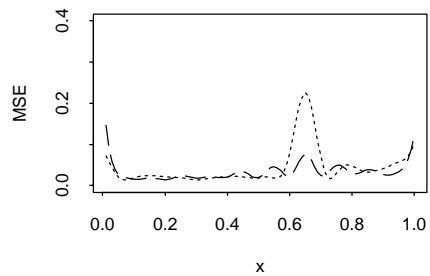
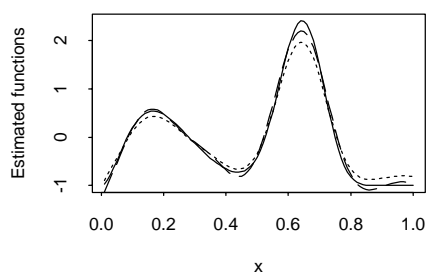
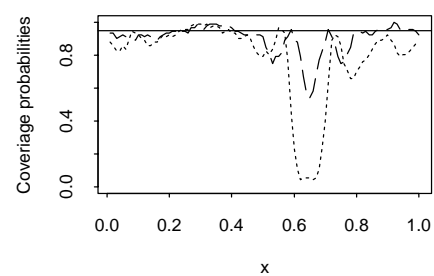
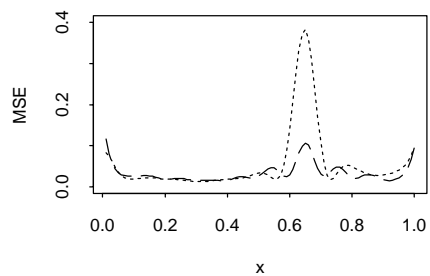
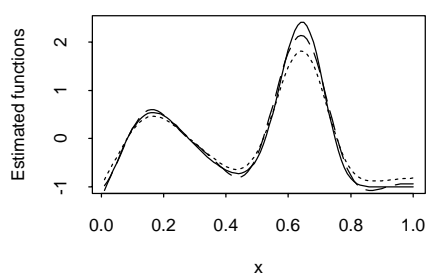


Figure 4(a)

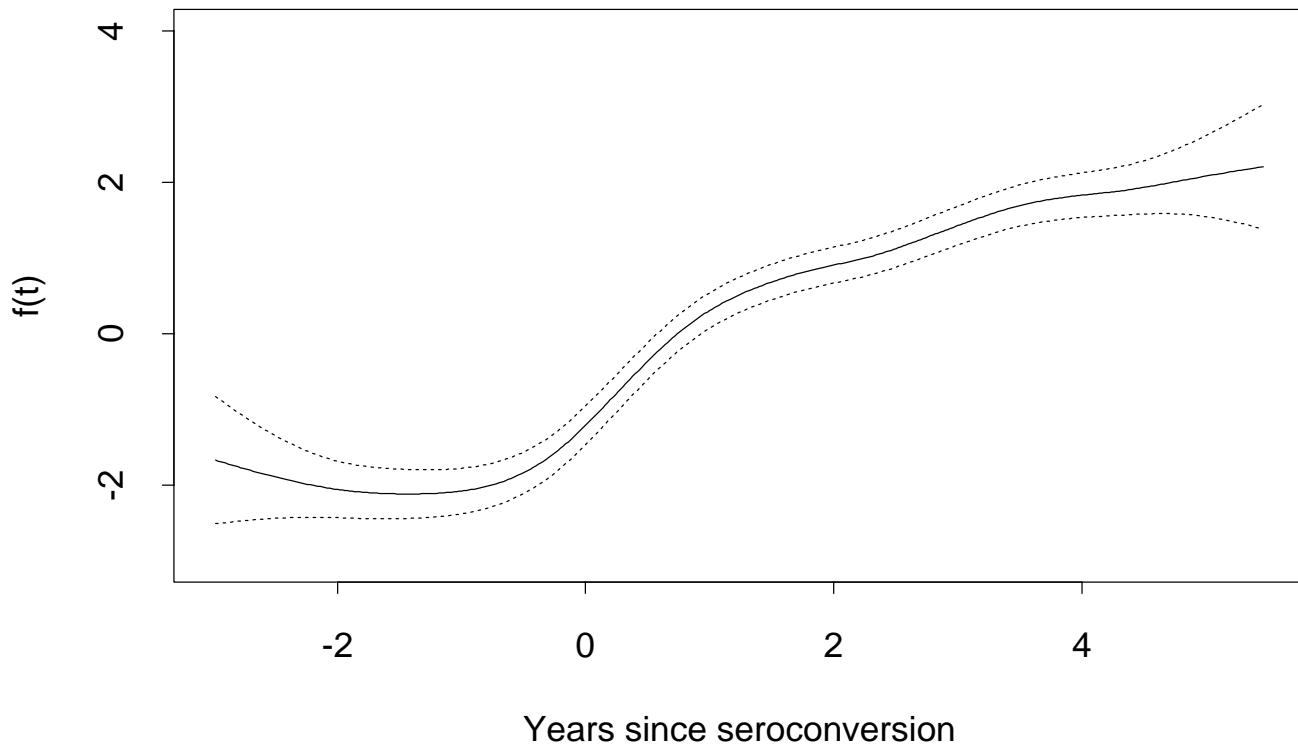


Figure 4(b)

