Muhammad Abu Shadeque Mullah and Andrea Benedetti*

Effect of Smoothing in Generalized Linear Mixed Models on the Estimation of Covariance Parameters for Longitudinal Data

DOI 10.1515/ijb-2015-0026

Abstract: Besides being mainly used for analyzing clustered or longitudinal data, generalized linear mixed models can also be used for smoothing via restricting changes in the fit at the knots in regression splines. The resulting models are usually called semiparametric mixed models (SPMMs). We investigate the effect of smoothing using SPMMs on the correlation and variance parameter estimates for serially correlated longitudinal normal, Poisson and binary data. Through simulations, we compare the performance of SPMMs to other simpler methods for estimating the nonlinear association such as fractional polynomials, and using a parametric nonlinear function. Simulation results suggest that, in general, the SPMMs recover the true curves very well and yield reasonable estimates of the correlation and variance parameters. However, for binary outcomes, SPMMs produce biased estimates of the variance parameters for high serially correlated data. We apply these methods to a dataset investigating the association between CD4 cell count and time since seroconversion for HIV infected men enrolled in the Multicenter AIDS Cohort Study.

Keywords: fractional polynomials, generalized linear mixed models, semiparametric mixed models, smoothing, splines

1 Introduction

The relationship between a quantitative risk factor and an outcome may take many different functional forms. To avoid the bias induced by misspecifying the functional form and the loss of efficiency in testing produced by categorizing continuous variables, nonparametric (flexible) regression models are often used to model the effects of continuous covariates [1, 2].

Generalized linear mixed models (GLMMs) [3], primarily used for analyzing overdispersed and correlated data (e.g. longitudinal data), can also be used for smoothing [4, 5]. To achieve a smooth function, we can use the GLMM to shrink the regression coefficients of knot points from a regression spline towards zero, by including them as random effects and constraining them to follow a normal distribution with mean zero and constant variance. This is equivalent to penalized splines and the resulting models are known as semiparametric mixed models (SPMMs) [6]. A key feature of this approach is that the smoothing parameter, which controls the trade-off between bias and variance may be directly estimated from the data [4]. Moreover, it is easily implemented in standard statistical software. Within a single model, we can estimate nonlinear covariate effects by penalized splines, while accommodating overdispersion and correlation by adding random effects to the additive predictor.

While the SPMM is evidently a sophisticated and useful method for smoothing, very limited work has been done to explore how well the covariance parameters are estimated when it is used for curve fitting in analyzing correlated data, especially longitudinal data (see, e.g., Lin and Zhang [7] and Chen et al. [8]). Longitudinal data consist of repeated measurements on one or more groups of individuals taken over time.

***Corresponding author: Andrea Benedetti,** Departments of Medicine and of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, E-mail: andrea.benedetti@mcgill.ca
Muhammad Abu Shadeque Mullah, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

The sequential nature of the measures implies that certain types of correlation structures are likely to arise. Serial correlation in the within-subject error structure is commonly modeled by using a first-order autoregressive, AR(1), model.

The objective of the present study is to investigate the effect of smoothing on covariance parameter estimation when analyzing overdispersed and serially correlated data using the SPMM to model both smooth curves and correlation at the same time. We compare smoothing of a single continuous covariate in a GLMM via restricting the knots in a regression spline to other simpler methods for estimating the nonlinear dose-response association such as fractional polynomials (FP), and a parametric nonlinear function of the covariate effect (e.g., quadratic) in data with correlated responses. We apply these methods to model the association between CD4 cell count and time since seroconversion for HIV infected men enrolled in the Multicenter AIDS Cohort Study [9].

Section 2 contains a brief description of the methods used for smoothing in correlated data. In section 3, we describe data generation and results from simulation studies. Section 4 provides application of smoothing methods to a dataset. The article concludes with a discussion in section 5.

2 Smoothing methods for correlated data

Let Y_{ij} be the response and $\left(\mathbf{w}_{ij}, x_{ij}, \mathbf{z}_{ij}\right)$ denote covariates measured at the j^{th} time (member) on the i^{th} subject (cluster), where $j=1,2,\ldots,n_i$ and $i=1,2,\ldots,m$. Here, $\mathbf{w}_{ij}$ is a covariate vector associated with fixed effects of a linear form, x_{ij} is a covariate associated with fixed effect of nonlinear functional form, and $\mathbf{z}_{ij}$ is a covariate vector associated with random effects. For simplicity, we consider smoothing only a single continuous covariate. Given the vector $\boldsymbol{b}_i$ of random effects, Y_{ij} are assumed to be observations from a distribution in the canonical exponential family

$$f_{Y_{ij}|\mathbf{b}_i}(y_{ij}|\boldsymbol{b}_i) = \exp\left[\left\{y_{ij}\eta_{ij}\left(\mathbf{w}_{ij}, x_{ij}, \mathbf{z}_{ij}\right) - \psi\left\{\eta_{ij}\left(\mathbf{w}_{ij}, x_{ij}, \mathbf{z}_{ij}\right)\right\}\right\}/a(\phi) + c(y_{ij};\phi)\right]$$

for known functions $a(\cdot)$, $\psi(\cdot)$ and $c(\cdot)$, and dispersion parameter ϕ, where η_{ij} is the canonical parameter. We consider a flexible generalized linear mixed model, where the regression function $\mu_{ij} = \mathbb{E}\left[Y_{ij}|\boldsymbol{b}_i\right] = \psi'\left\{\eta_{ij}\left(\mathbf{w}_{ij}, x_{ij}, \mathbf{z}_{ij}\right)\right\}$ is modeled via a link function g,

$$g\left(\mathbb{E}\left[Y_{ij}|\boldsymbol{b}_i\right]\right) = \mathbf{w}_{ij}^T\gamma + f(x_{ij}) + \mathbf{z}_{ij}^T\boldsymbol{b}_i, \tag{1}$$

where γ is vector of regression coefficients, and $f(\cdot)$ is an arbitrary smooth function. Note that $\mathbf{w}_{ij}$ does not include an intercept as it is subsumed in the function for x_{ij}. The random effects $\boldsymbol{b}_i$ are assumed to be independently distributed as $N(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. In addition to the correlation induced by the random effects, repeated responses are assumed to be serially correlated conditional on $\boldsymbol{b}_i$ such that

$$\text{Corr}\left[Y_{ij}, Y_{ij'}|\boldsymbol{b}_i\right] = \rho_{jj'} \quad \text{for all } (j, j').$$

One can choose an autocorrelation model to determine the $\rho_{jj'}$. Common choices include spatial exponential or spatial Gaussian models, an AR(1) model, etc. In many longitudinal studies the response is modeled as a nonlinear function of time for each individual (see, e.g., Zeger and Diggle [10] and Zhang et al. [11]). However, other covariates (time dependent or baseline) may also be nonlinearly associated with outcomes.

For a linear function, $f(\cdot)$ in (1) simply takes the form $f(x) = \beta_0 + \beta_1 x$. The linearity assumption may not always be appropriate. There are many different ways to estimate the smooth function $f(\cdot)$ (see, Rice and Wu [12] and Guo [13], among others). However, this study considers the following methods to estimate smooth dose response curves.

2.1 Parametric nonlinear function

By parametric nonlinear function we refer to a regression model whose mean is linear in parameters but has higher order of covariates. The simplest and most common way to represent curvature in regression models is using polynomials of the covariates, typically quadratics. Adding a parametric nonlinear term (e.g., quadratic), $f(\cdot)$ in (1) can be modeled as

$$f(x)=\beta_0+\beta_1 x+\beta_2 x^2. \tag{2}$$

One advantage of this parametric approach is that it is more efficient if the model is correct. However, conventional low order polynomials do not always fit the data very well. High order polynomials follow the data more closely but often fit poorly at extreme values of the covariates [14].

2.2 Fractional polynomials

Fractional polynomials (FP), proposed by Royston and Altman [14], are a family of curves whose power terms are restricted to a small predefined set of integer and non-integer values. For one covariate x, a FP of degree d with powers $p_1, \ldots, p_d$ is given by

$$FP_d(x)=\beta_0+\beta_1 x^{p_1}+\ldots+\beta_d x^{p_d}, \tag{3}$$

where powers $p_1, \ldots, p_d$ are taken from S,

$$S=\{-2,-1,-0.5,0,0.5,1,2,3,\ldots,\max(3,d)\}.$$

Here, x^0 is defined as $log_e x$. Usually, $d=1$ or $d=2$ is sufficient for a good fit as models with degree higher than two are rarely required in practice [14]. Ambler and Royston [15] suggested an iterative algorithm for covariate selection and model fitting when several covariates are available. For $d=2$, their algorithm first selects the best-fitting second-degree FP which has the lowest deviance (-2 × log likelihood among all possible models with two powers, $(p_1,p_2)\in S$. The best selected model is then tested against the null model, a straight line, and the best-fitting first-degree FP (with lowest deviance among all models with power $p_1\in S$) in an attempt at further simplification. All significance tests are performed by calculating an approximate P-value based on the difference in deviances having a chi-squared or F distribution, depending on the regression in use. This approach has been implemented in most popular statistical software (e.g. the mfp package in R, macros in SAS (www.imbi.uni-freiburg.de/biom/mfp), and the fp function in Stata).

2.3 Semiparametric mixed models (SPMMs)

The nonlinear association between an outcome and covariates can be modeled using penalized splines within the framework of a mixed effects model. This approach is a trade-off between regression splines (which rely on the number and position of knots) and smoothing splines (which require intensive computation for larger datasets) [16]. To capture the structure of $f(\cdot)$ in (1), we define K distinct knots $t_1, \ldots, t_K$ in the range of x and consider the regression spline

$$f(x)=\beta_0+\beta_1 x+\ldots+\beta_p x^p+\sum_{k=1}^{K}u_k(x-t_k)_+^p, \tag{4}$$

where $\beta_0, \ldots, \beta_p$ are the regression coefficients, u_k denotes the spline coefficient at knot t_k, p is the degree of polynomial used (e.g., two for quadratic, three for cubic), and the truncated basis function component $(x-t_k)_+^p=\max\{0,(x-t_k)^p\}$. Note that when $p=1$ the knot coefficients u_k represent changes in slope from one segment to the next. Unpenalized estimation of u_k would lead to a bumpy fit due to the large number of truncated polynomials. To avoid overfitting we assume the u_k are independently distributed as

$$u_k \sim \mathrm{N}(0, \sigma_u^2). \tag{5}$$

This constraint shrinks the u_k towards zero to reduce the magnitude of slope changes, leading to a smoother fit.

Denoting $\mathbf{w} = [\mathbf{w}_{11} \ \ldots \ \mathbf{w}_{mn_m}]^T$, $\mathbf{x} = [\mathbf{x}_{11} \ \ldots \ \mathbf{x}_{mn_m}]^T$, $\mathbf{z} = [\mathbf{z}_1 \ \ldots \ \mathbf{z}_m]^T$ with $\mathbf{z}_i = [\mathbf{z}_{i1} \ \ldots \ \mathbf{z}_{in_i}]^T$, the vector of fixed effects $\boldsymbol{\beta} = (\gamma^T, \beta_0, \ldots, \beta_p)^T$, vector of random effects $\mathbf{u} = (u_1, \ldots, u_K, b_1, \ldots, b_m)^T$, the associated design matrices $\mathbf{X} = \left[\mathbf{w} \quad \mathbf{1} \quad \mathbf{x} \quad \mathbf{x}^2 \ \ldots \ \mathbf{x}^p\right]$ and $\mathbf{Z} = \left[(\mathbf{x} - t_1 1)_+^p \ \ldots \ (\mathbf{x} - t_K 1)_+^p \quad \underset{1 \le i \le m}{\text{blockdiagonal}} \ \mathbf{z}_i\right]$, where 1 is the vector of all ones, and the response vector $\mathbf{Y} = [Y_{11} \ \ldots \ Y_{mn_m}]^T$, we can rewrite (1), (4) and (5) as the generalized linear mixed model

$$g(\mathbb{E}[\mathbf{Y}|\mathbf{u}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \tag{6}$$

where

$$\mathrm{Cov}(\mathbf{u}) = \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \underset{1 \le i \le m}{\text{blockdiagonal}} \ \Sigma(\boldsymbol{\theta}) \end{bmatrix}.$$

This mixed model representation of penalized splines allows us to take full advantage of existing methods and software for GLMMs. The resulting models are called semiparametric mixed models (SPMMs) [6]. We discuss maximum likelihood (ML) estimation of the SPMMs in the subsequent subsection.

An important special case of (6) is the Gaussian mixed model with random intercepts $b_i \sim \mathrm{N}(0, \sigma_b^2)$. Fitting such a SPMM is equivalent to minimizing the penalized least squares (PLS)

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \lambda \sum_{k=1}^{K} u_k^2 + \frac{\sigma_\varepsilon^2}{\sigma_b^2} \sum_{i=1}^{m} b_i^2,$$

where the smoothing parameter $\lambda = \sigma_\varepsilon^2 / \sigma_u^2$. Here, σ_ε^2 is the variance of the random error, and for a general vector $\mathbf{v}$, $\|\mathbf{v}\| \equiv \sqrt{\mathbf{v}^\mathbf{T} v}$ is the usual Euclidean norm of the vector $\mathbf{v}$. One could take the approach of minimizing the PLS directly, but obtaining ML estimates from the SPMM is advantageous because the smoothing parameter, λ, is estimated directly from the data as $\hat{\lambda} = \hat{\sigma}_\varepsilon^2 / \hat{\sigma}_u^2$, though it is possible to specify it directly if desired [4]. Moreover, using the SPMM approach to estimate the smoothing parameter works better (in terms of mean squared error performance and computational stability) than generalized cross validation (GCV), Akaike information criterion (AIC), etc. [17].

SPMMs require the number and position of the knots to be specified in advance in addition to specifying the degree (p) of polynomial for the curve segments. In practice, the simplest choice for p is 1 (see, for example, Gurrin et al. [4], Wand [5] and Durban et al. [16]) so that the truncated line basis for the knots $t_1, \ldots, t_K$ is

$$1, x_{ij}, (x_{ij} - t_1)_+, \ldots, (x_{ij} - t_K)_+.$$

These bases are preferred because of their simple mathematical form, which is very useful when formulating complicated models [5]. Greater smoothness can be achieved using higher degree truncated polynomial bases such as truncated cubic ($p = 3$) basis functions, however the use of truncated polynomial ($p \ge 2$) bases has been criticized by many authors because of their sub-optimal numerical properties [5] and poor behaviour in the tails [18]. One solution to this is to use natural cubic splines. More complex bases such as B-splines [19] or radial basis functions [6] that have better numerical properties could also be easily used to fit models under this framework.

2.4 Estimation

Once f is specified by any of the methods discussed above, model (1) can be written in a general form

$$g\left(\mathbb{E}\left[\mathbf{Y}|\mathbf{u}^{\star}\right]\right)=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}^{\star},$$

where $\mathbf{u}^{\star}\sim f_{u^{\star}}(\mathbf{u}^{\star}|\mathbf{G})=N(\mathbf{0},\mathbf{G})$. Let $\boldsymbol{\rho}$ be the vector of components in $\operatorname{Corr}\left[Y_{ij},Y_{ij'}|\mathbf{u}_i^{\star}\right]$. The likelihood function is then given by

$$L(\boldsymbol{\beta},\phi,\mathbf{G},\boldsymbol{\rho}|y_i)=\prod_{i=1}^{m}\int f(y_i|\mathbf{u}_i^{\star},\boldsymbol{\beta},\phi,\boldsymbol{\rho})f(\mathbf{u}_i^{\star}|\mathbf{G})\;d\mathbf{u}_i^{\star}. \tag{7}$$

Except in the case of normally distributed outcomes, the likelihood in (7) does not have a closed form as the integral is intractable. In these cases, numerical approximations are required. While various approximation procedures and Bayesian methods using Gibbs sampling and EM algorithm have been proposed in the literature (see, Demidenko [20] for details), the penalized quasi-likelihood (PQL) method of Breslow and Clayton [3] is the most frequently used estimation technique because of its simplicity and relatively small computational effort. Lin and Zhang [7] proposed the double penalized quasi-likelihood to make approximate inference in generalized additive mixed models using smoothing splines. For the linear mixed effects model, obtaining maximum likelihood estimates from (7) is quite straightforward. However, to obtain less biased estimates for the covariance parameters, restricted maximum likelihood estimation (REML) is often used.

3 Simulation studies

To empirically and systematically compare smoothing approaches according to the effect of smoothing on correlation and variance parameters, we performed a series of simulation experiments with data simulated from and analyzed using models with correlated errors and a subject-specific random effect.

3.1 Data generation

3.1.1 Overview

We generated longitudinal-type data such that the repeated responses were serially correlated. We considered an extended model, where in addition to the serial correlation, repeated responses were assumed to be influenced by an individual random effect causing them to be overdispersed. Responses were generated from normal, Poisson and binary distributions. For each distribution of the response variable, we generated data while varying: homogeneous or heterogeneous cluster sizes (n_i), number of clusters (m), magnitude of the serial correlation coefficient (ρ), and the form of the association between response and covariate(s). See Table 1 for the values used in data generation. Not every combination of data generation parameters was used. We chose a range of plausible values for each parameter that might be encountered in everyday data analysis. For each experiment, 1,000 datasets were simulated.

3.1.2 Response models for simulation

Repeated observations y_{ij} were generated within each subject (cluster) according to the model

$$g\left(\mathbb{E}\left[Y_{ij}|b_i\right]\right)=\beta x_{1i}+f(x_{2ij})+b_i, \tag{8}$$

Table 1: Parameter values used for data generation.[a,b,c]

Items	Values
Number of clusters, m	100, 500
Cluster size, n_i	5, 1–10
AR(1) correlation coefficient, ρ	
Normal and Poisson responses	0, 0.25, 0.5, 0.75
Binary response	0, 0.25, 0.5
Random effects variance, σ_b^2	
Normal response	0, 0.1
Poisson and Binary responses	0, 0.25
Random error variance, σ_ε^2	
Normal response	0.1

Note: [a]For normal data, we considered the following combinations of $(\rho, \sigma_b^2, \sigma_\varepsilon^2)$: $(0, 0.1, 0.1)$; $(0.25, 0.1, 0.1)$; $(0.5, 0.1, 0.1)$; $(0.75, 0.1, 0.1)$; $(0.75, 0, 0.1)$. [b]For Poisson data, the combinations of (ρ, σ_b^2) were: $(0, 0.25)$; $(0.25, 0.25)$; $(0.5, 0.25)$; $(0.75, 0.25)$; $(0.75, 0)$. [c]For binary data, the combinations of (ρ, σ_b^2) were: $(0, 0.25)$; $(0.25, 0.25)$; $(0.5, 0.25)$; $(0.5, 0)$.

where b_i are the independent subject-specific random intercepts from $N(0, \sigma_b^2)$, $\beta = 0.5$, x_{1i} takes value 1 for half of the subjects and 0 for the others, mimicking a group membership indicator or a binary treatment indicator, x_{2ij} is in $[0, 1]$, and $f(x_{2ij})$ is assumed to be one of the following three test functions:

$$f_1(x) = x^{11}[10(1-x)]^{4.5} + 10(10x)^3(1-x)^{10} - 3.396, \tag{9}$$

$$f_2(x) = (x - Q_1(x))^2 * I(x \geq Q_1(x)) + 0.5, \tag{10}$$

$$f_3(x) = x, \tag{11}$$

where in (10), $Q_1(x)$ is the first quartile of x and $I(\cdot)$ is an indicator function. All three functions are plotted in the left column of Figure 1. The shapes of the functions in (9), (10) and (11) are referred to later as double hump (DH), linear quadratic threshold at 25th percentile (QT), and linear (LIN), respectively. We considered x_1 and x_2 to be independent.

In the normal and binary cases, we assumed the covariate x_2 to be varying within each cluster taking m values equally spaced in $[0, 1]$. Specifically, we considered

$$x_{2ij} = \frac{\text{floor}\{(i + n_i - 1)/n_i\}}{m} + \frac{(j-1)}{n_i}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where floor$(\cdot)$ denotes a truncated operator. Here, x_2 mimics the "time" or "age" covariate in a longitudinal data settings. For Poisson distributed data, the correlation structure for the repeated counts is affected by time dependent covariates [21]. This may lead to a non-stationary or weakly identified autocorrelation structure in the presence of a time-varying covariate. Therefore, for simplicity, and to ensure a stationary autocorrelation structure, we considered x_2 to be a cluster level (time independent, e.g., age at base line) covariate taking m distinct values equally spaced between 0 and 1. In particular, for Poisson data, we let

$$x_{2ij} = x_{2i} = i/m.$$

For each distribution, true response means were obtained by using the inverse of the canonical link to the linear predictor, and data were simulated from the appropriate distribution with that mean. Note that to obtain the true linear predictor in (8), we scaled each function $f_1(x), f_2(x), f_3(x)$ with respect to x_2 as follows: (a) in the normal case the functions were scaled to have range $[0, 1]$; (b) in the Poisson case the functions were scaled so that the true means lay in the range $[1, 8]$; (c) in the binary case the functions were scaled so that the success probabilities lay in $[0.25, 0.75]$. These ranges of means and probabilities did not include the effects of the fixed effect βx_{1i} and random effect b_i.

3.1.3 Outcome variables

Conditional on the random effect b_i, serially correlated responses were generated from all data distributions according to an AR(1) model for the correlation structure so that

$$\mathrm{Corr}\left[Y_{ij}, Y_{ij'}|b_i\right] = \rho^{|j'-j|}, \quad \text{for all } (j, j').$$

Normal Responses.

We generated correlated normal responses according to model (8) with identity link

$$Y_{ij} = \beta x_{1i} + f(x_{2ij}) + b_i + \epsilon_{ij}. \tag{12}$$

Conditional on the random effect b_i, we generated AR(1) correlated responses by considering

$$\epsilon_{ij} \sim \mathrm{MVN}_{n_i}\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma_\varepsilon^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_i-1} \\ \rho & 1 & \rho & \dots & \rho^{n_i-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \dots & 1 \end{pmatrix}\right].$$

Poisson Responses.

Correlated Poisson responses were generated from model (8) using the link $g(\mathbb{E}[Y_{ij}|b_i]) = \log(\mathbb{E}[Y_{ij}|b_i])$. Given the random effect b_i, stationary autocorrelated counts were simulated from AR(1) model following Sutradhar [21], Section 6.

Binary Responses.

To generate binary responses, we used model (8) with the link function $g(\mathbb{E}[Y_{ij}|b_i]) = \mathrm{logit}(\mathbb{P}[Y_{ij} = 1|b_i])$. Conditional on the random effect b_i, we adopted Qaqish's [22] multivariate binary model to generate AR(1) serial correlated responses.

3.2 Analysis of simulated data

Each simulated dataset was analysed by fitting a generalized linear mixed model that included a random intercept and AR(1) serial correlation structure. In some scenarios we generated data with either the serial correlation (ρ) or the random effect variance (σ_b^2) set to zero, but fit models that included both random intercept and AR(1) correlation. We call these "misspecified" although they are just special cases with extreme values of the model parameters. Model parameters were estimated via restricted maximum likelihood estimation (REML) for normally distributed data and the penalized quasi likelihood (PQL) approach of Breslow and Clayton [3] for Poisson and binary data. In each case, we adopted the following approaches to estimate the smooth curve, $f(\cdot)$:

1. Semiparametric mixed model (SPMM) using truncated line bases with number and position of knots specified by the simple and reasonable default rule given by Wand [5] as

$$t_k = \left(\frac{k+1}{K+2}\right)\text{th sample quantile of unique } x\text{'s}, \tag{13}$$

where $1 \le k \le K$, and $K = \min(n/4, 35)$. For the parameters of all the fixed effects and the smooth term, we obtained the approximate (frequentist) covariance matrix following the approach described in Lin and Zhang [7].

2. Fractional polynomials (FP) of degree 2 to identify the best curve by ignoring the correlated nature of the data. To select the best FP, we followed an algorithm proposed by Ambler and Royston [15] (as implemented in mfp package in R) that has been described in Section 2.2. We then fit the GLMM to accommodate the correlation structure of the data using the best selected curve from the FP fit.
3. Quadratic polynomials as the simplest parametric non-linear function.
4. Linear function.

As a sensitivity analysis we also fit a model using the true data-generating curve. For the linear, quadratic polynomial and FP, curve $f(\cdot)$ was estimated as a fixed effects component. The approximate standard error of the estimated curve $\hat{f}(\cdot)$ was obtained conditional on the estimates of the random-effects parameters and therefore this only accounted for the uncertainty of the fixed effects. All simulations and analyses were carried out in R software employing lme and glmmPQL functions for REML and PQL methods, respectively. For both functions default procedure values were used unless otherwise noted. We allowed 200 iterations for convergence in all models fit.

3.3 Performance indicators

For the covariance parameters estimators $\hat{\rho}$, $\hat{\sigma}_b^2$, $\hat{\sigma}_\varepsilon^2$, and fixed effect estimator $\hat{\beta}$, we computed percentage relative biases (PRBs), simulated mean squared errors (MSEs), and empirical coverage probabilities (CPs). PRB was computed as

$$\text{PRB} = \frac{\text{Bias}}{\text{True Value}} \times 100$$

while the empirical coverage probability (CP) of an estimator $\hat{\alpha}$ of α was obtained as

$$\text{CP}(\hat{\alpha}) = \frac{1}{1000} \sum_{r=1}^{1000} I(\hat{\alpha}_{Lr} \le \alpha \le \hat{\alpha}_{Ur}),$$

where $\hat{\alpha} \in \{\hat{\rho}, \hat{\sigma}_b^2, \hat{\sigma}_\varepsilon^2, \hat{\beta}\}$, and $\hat{\alpha}_L$ and $\hat{\alpha}_U$, respectively, are the lower and upper limits of the approximate confidence intervals for α. The approximate confidence intervals for the parameters were obtained using a normal approximation to the distribution of the (restricted) maximum likelihood estimators following Pinheiro and Bates [23].

For the smoothing curve estimators, $\hat{f}$, we computed pointwise mean average squared distance/error (MASE) from the true curve, and the mean average coverage probability (MACP). The pointwise MASE was defined as the mean over the 1,000 replicated datasets of the pointwise average squared error,

$$\text{ASE} = \left(\sum_{i=1}^{m} n_i\right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \{\hat{f}(x_{ij}) - f(x_{ij})\}^2.$$

The 95% pointwise MACP was obtained as the mean of the 1,000 pointwise average coverage probabilities,

$$\text{ACP} = \left(\sum_{i=1}^{m} n_i\right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I\left(|\hat{f}(x_{ij}) - f(x_{ij})| \le 1.96 \times SE(\hat{f}(x_{ij}))\right),$$

where $SE(\hat{f}(x_{ij}))$ is the standard error of $\hat{f}(x_{ij})$.

To compare the fits graphically, we plotted the mean fitted values and 95% pointwise coverage probabilities of the true function. At each observed value of x, the mean fitted value was obtained by taking the average over the 1,000 replications. Pointwise coverage probability was computed as the proportion of replications in which the confidence interval contained the true curve.

3.4 Simulation results

We analyzed the simulated data for 100 and 500 clusters with homogeneous ($n_i = 5$) and heterogeneous ($1 \le n_i \le 10$) cluster sizes. Since the results showed consistent patterns, we report only the case of $n_i = 5$ for $m = 500$. Results from one scenario with varying cluster sizes ($1 \le n_i \le 10$) are also presented in the supplementary materials (Supplementary Table 4).

Simulation results for normally distributed data reported in Table 2 indicate that for complex functional forms (e.g., double hump), the SPMM method yielded very good estimates for the correlation, variance, and fixed effect parameters in terms of bias and MSE. It also performed very well in estimating the nonparametric function (with respect to MASE) and had the correct average coverage probability. On the contrary, FP, quadratic, and linear fits performed very poorly in estimating all parameters except for the fixed effect (β). For slightly nonlinear associations (e.g., linear quadratic threshold), the SPMM approach estimated all the parameters extremely well and produced very good coverage probability for the nonparametric

Table 2: For normally distributed outcomes, Percentage Relative Bias (PRB) and Mean Squared Error (MSE) of the estimates for ρ, σ_b^2, σ_ε^2 and β, and Mean Average Squared Error (MASE) with Mean Average Coverage Probability (MACP) of the estimated curve for f obtained from models fit using different smoothing methods.

						Outcome Distribution: Normal Number of Cluster = 500, Cluster Size = 5
Shape	**True Value**	**Quantity**				**Smoothing Methods**
			Linear	**Quadratic**	**FP**	**SPMM**
Double Hump	$\rho = 0.5$	PRB	−100.00	−99.43	−74.38	0.16
		MSE	0.2500	0.2472	0.1393	0.0012
	$\sigma_b^2 = 0.1$	PRB	17.73	20.36	17.90	−0.15
		MSE	0.0004	0.0005	0.0004	0.0001
	$\sigma_\varepsilon^2 = 0.1$	PRB	45.48	31.93	14.21	0.44
		MSE	0.0021	0.0010	0.0002	0.0000
	$\beta = 0.5$	PRB	−0.05	−0.05	−0.04	−0.05
		MSE	0.0012	0.0012	0.0012	0.0012
	$f = f_1$	MASE	0.0633	0.0524	0.0325	0.0010
		MACP	0.21	0.25	0.21	0.94
Quadratic Threshold	$\rho = 0.5$	PRB	10.76	−0.19	−0.19	0.14
		MSE	0.0037	0.0011	0.0011	0.0011
	$\sigma_b^2 = 0.1$	PRB	−20.5401	0.0362	−0.0012	−0.0912
		MSE	0.0005	0.0001	0.0001	0.0001
	$\sigma_\varepsilon^2 = 0.1$	PRB	36.73	0.56	0.53	0.49
		MSE	0.0014	0.0000	0.0000	0.0000
	$\beta = 0.5$	PRB	−0.06	−0.06	−0.05	−0.05
		MSE	0.0012	0.0012	0.0012	0.0012
	$f = f_2$	MASE	0.0168	0.0008	0.0007	0.0007
		MACP	0.20	0.93	0.93	0.96
Linear	$\rho = 0.5$	PRB	0.09	0.10	0.09	0.10
		MSE	0.0011	0.0011	0.0011	0.0011
	$\sigma_b^2 = 0.1$	PRB	−0.0704	−0.0770	−0.0752	−0.0680
		MSE	0.0001	0.0001	0.0001	0.0001
	$\sigma_\varepsilon^2 = 0.1$	PRB	0.4949	0.5051	0.4952	0.4635
		MSE	0.0000	0.0000	0.0000	0.0000
	$\beta = 0.5$	PRB	−0.0541	−0.0539	−0.0536	−0.0464
		MSE	0.0012	0.0012	0.0012	0.0012
	$f = f_3$	MASE	0.0006	0.0006	0.0006	0.0006
		MACP	0.95	0.95	0.95	0.95

function. The FP and quadratic fits performed similarly to SPMM although they had slightly lower coverage probabilities for nonparametric function estimates. Using a linear functional form, however, performed poorly. When the true association was linear, results from all the methods were very similar with almost unbiased estimates for all parameters and small MSEs. The mean coverage probabilities for the estimated function were exactly 95% for all methods. Similar results were observed when $\rho = 0.75, \sigma_b^2 = \sigma_\varepsilon^2 = 0.1$ and $\rho = 0.25, \sigma_b^2 = \sigma_\varepsilon^2 = 0.1$ (data not shown).

In the two slightly misspecified cases, that is, when the correlation was due to (i) only random intercepts but no serial correlation ($\rho = 0, \sigma_b^2 = \sigma_\varepsilon^2 = 0.1$) or (ii) serial correlation but no random intercepts ($\rho = 0.75, \sigma_b^2 = 0, \sigma_\varepsilon^2 = 0.1$), the results were similar to those that have been reported above except that the estimates of σ_ε^2 were slightly biased in case (ii) for all smoothing methods and shapes (see supplementary Table 1).

Table 3 shows results when the responses were Poisson distributed. The SPMM approach worked well in estimating all parameters and had average coverage probability for $f(x_2)$ close to the nominal value when the true functional form was complex (e.g., double hump). FP, quadratic, and linear fits, on the other hand, performed badly in estimating the variance of the random effects and $f(x_2)$, although the auto-correlation and fixed effect (β) were reasonably estimated. For the quadratic threshold shape, SPMM, FP, and quadratic fits estimated all the parameters reasonably well. As expected, the linear fit estimated the random effects variance and $f(x_2)$ poorly. For the linear functional form, all methods performed similarly and produced good estimates for all parameters. Similar results were observed for all other scenarios investigated:

Table 3: For Poisson distributed outcomes, Percentage Relative Bias (PRB) and Mean Squared Error (MSE) of the estimates for ρ, σ_b^2 and β, and Mean Average Squared Error (MASE) with Mean Average Coverage Probability (MACP) of the estimated curve for f obtained from models fit using different smoothing methods.

						Outcome Distribution: Poisson Number of Cluster = 500, Cluster Size = 5
Shape	**True Value**	**Quantity**	**Smoothing Methods**			
			Linear	**Quadratic**	**FP**	**SPMM**
Double Hump	$\rho = 0.5$	PRB	−4.81	−4.27	−1.02	−1.02
		MSE	0.0012	0.0014	0.0044	0.0015
	$\sigma_b^2 = 0.25$	PRB	94.82	80.63	41.99	−0.24
		MSE	0.0578	0.0421	0.0163	0.0007
	$\beta = 0.5$	PRB	−3.88	−3.64	−3.26	−2.83
		MSE	0.0055	0.0052	0.0047	0.0035
	$f = f_1$	MASE	0.2438	0.2035	0.1445	0.0120
		MACP	0.27	0.33	0.33	0.93
Quadratic Threshold	$\rho = 0.5$	PRB	−6.44	−6.34	−6.41	−6.35
		MSE	0.0015	0.0018	0.0018	0.0018
	$\sigma_b^2 = 0.25$	PRB	21.44	1.84	1.41	1.99
		MSE	0.0040	0.0009	0.0009	0.0009
	$\beta = 0.5$	PRB	−2.64	−2.57	−2.59	−2.60
		MSE	0.0046	0.0041	0.0041	0.0041
	$f = f_2$	MASE	0.0602	0.0067	0.0083	0.0073
		MACP	0.23	0.91	0.87	0.92
Linear	$\rho = 0.5$	PRB	−3.40	−3.42	−3.47	−3.30
		MSE	0.0011	0.0011	0.0011	0.0011
	$\sigma_b^2 = 0.25$	PRB	−2.19	−2.45	−2.71	−2.30
		MSE	0.0006	0.0006	0.0006	0.0006
	$\beta = 0.5$	PRB	−2.47	−2.48	−2.47	−2.54
		MSE	0.0033	0.0033	0.0033	0.0033
	$f = f_3$	MASE	0.0042	0.0050	0.0072	0.0043
		MACP	0.94	0.95	0.90	0.94

Table 4: For binary distributed outcomes, Percentage Relative Bias (PRB) and Mean Squared Error (MSE) of the estimates for ρ, σ_b^2 and β, and Mean Average Squared Error (MASE) with Mean Average Coverage Probability (MACP) of the estimated curve for f obtained from models fit using different smoothing methods.

Outcome Distribution: Binary **Number of Cluster = 500, Cluster Size = 5**						
Shape	**True Value**	**Quantity**	**Smoothing Methods**			
			Linear	**Quadratic**	**FP**	**SPMM**
Double Hump	$\rho = 0.25$	PRB	−14.28	−13.04	−8.10	−0.40
		MSE	0.0022	0.0020	0.0014	0.0011
	$\sigma_b^2 = 0.25$	PRB	4.52	4.42	3.49	1.20
		MSE	0.0167	0.0166	0.0193	0.0286
	$\beta = 0.5$	PRB	−7.51	−6.59	−5.16	−3.54
		MSE	0.0123	0.0122	0.0120	0.0120
	$f = f_1$	MASE	0.1886	0.1590	0.1042	0.0178
		MACP	0.40	0.41	0.45	0.92
Quadratic Threshold	$\rho = 0.25$	PRB	−2.76	−2.72	−2.38	−2.12
		MSE	0.0010	0.0010	0.0010	0.0010
	$\sigma_b^2 = 0.25$	PRB	−11.07	9.38	8.46	7.00
		MSE	0.0221	0.0247	0.0244	0.0240
	$\beta = 0.5$	PRB	−4.76	−3.48	−3.50	−3.76
		MSE	0.0123	0.0124	0.0124	0.0124
	$f = f_2$	MASE	0.0506	0.0106	0.0133	0.0125
		MACP	0.46	0.94	0.90	0.92
Linear	$\rho = 0.25$	PRB	−3.10	−3.21	−3.11	−3.16
		MSE	0.0010	0.0010	0.0010	0.0010
	$\sigma_b^2 = 0.25$	PRB	8.08	8.86	8.16	8.66
		MSE	0.0222	0.0224	0.0223	0.0220
	$\beta = 0.5$	PRB	−3.98	−3.93	−3.97	−3.78
		MSE	0.0123	0.0123	0.0123	0.0123
	$f = f_3$	MASE	0.0082	0.0099	0.0096	0.0084
		MACP	0.94	0.95	0.94	0.94

$\rho = 0.75, \sigma_b^2 = 0.25$; $\rho = 0.25, \sigma_b^2 = 0.25$; $\rho = 0.75, \sigma_b^2 = 0$; $\rho = 0, \sigma_b^2 = 0.25$ (see supplementary Table 2 for the results from last two scenarios).

Table 4 presents the simulation results for binary response data. Results suggest that for the double hump shape, the SPMM technique outperformed all other methods and estimated all the parameters fairly well. For the linear quadratic threshold and linear shapes of association, the SPMM, FP, and quadratic approaches estimated all parameters well except that the σ_b^2 estimates were positively biased. As ρ increased, the bias in estimating σ_b^2 also increased for all shapes of association (see supplementary Table 3). The estimates of σ_b^2 were found to be biased for all methods in the slightly misspecified cases, i.e., when $\rho = 0, \sigma_b^2 = 0.25$ and $\rho = 0.5, \sigma_b^2 = 0$ (supplementary Table 3). While the results are not shown, we note that biases similar to those given by the SPMM method existed even when the true data-generating model was used for estimation.

The estimated (model based) standard errors (SEs) of the parameter estimates ($\hat{\beta}, \hat{\rho}, \hat{\sigma}_b^2$ and $\hat{\sigma}_\varepsilon^2$) agreed well with the empirical (simulation based) standard errors for all methods in all cases except for $\hat{\sigma}_b^2$ in the binary data, where empirical SEs were 3% to 28% larger than estimated SEs for all methods (results not shown).

The coverage probability (CP) was generally near nominal level for β for all methods in all cases. For ρ, σ_b^2 and σ_ε^2, the CPs obtained from the SPMM approach were close to the nominal level for all shapes of association considered while for other methods CPs varied depending of the true shape of the association (results not shown).

The left panel of Figure 1 illustrates the ability of the four estimation methods to recover the true functions by comparing the four estimated curves based on 1,000 replications for normally distributed responses. The SPMM reconstructed the true curves extremely well for all the shapes. FP and quadratic fits performed well in capturing the true curves for linear quadratic threshold and linear shapes, whereas both the methods yielded unsatisfactory fits for the double hump function.

The right panel of Figure 1 compares the empirical pointwise coverage probabilities of the 95% confidence intervals of three functions calculated using four different methods. For normal data, the coverage probabilities of all confidence intervals were very close to 95% at well estimated values of x_2. For the double hump curve, the SPMM gave undercoverage for small values of $x_2 \leq 0.1$, i.e., the region where the signal-to-noise ratio was low. Afterwards, the coverage probability remained close to 95%. For quadratic threshold shape, the coverage probability of the SPMM confidence interval agreed slightly better with the nominal value throughout the range of x_2 and performed better than other methods. When the true functional form was linear, confidence intervals from all methods provided correct coverage probabilities although SPMM had slightly inferior performance.

Similar results in estimating nonparametric functions were obtained for Poisson and binary responses. However, the SPMM estimated curves were slightly biased and had slightly lower coverage probabilities (see supplementary Figures 1 and 2).

Convergence or singularity problems for the SPMM method were relatively minor (< 0.5%) and only occurred for the case when the true functional form was linear. Results from cases that did not converge were omitted.

4 Application to CD4 cell data

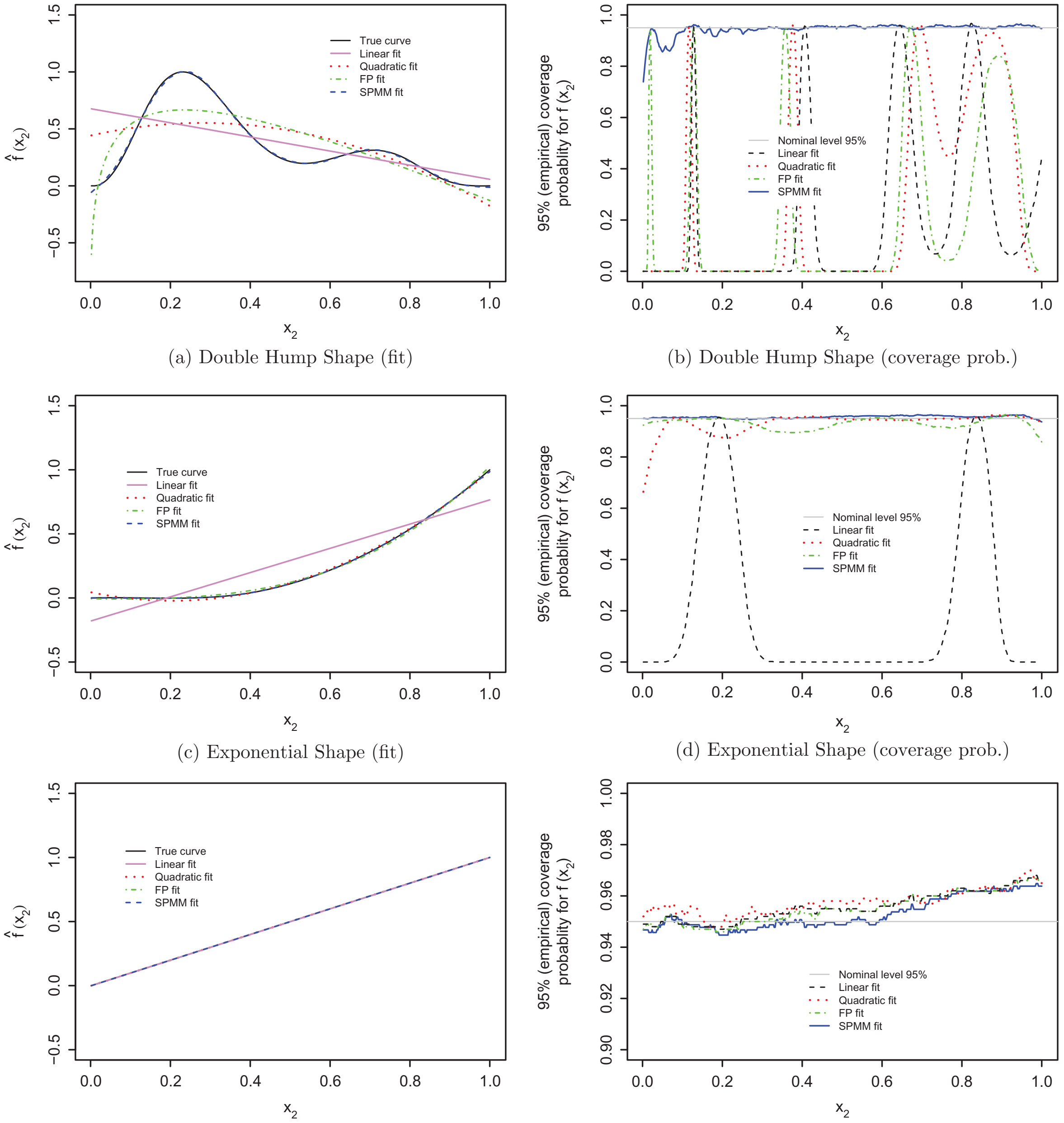
In this section we used both linear mixed models (LMMs) and generalized linear mixed models (GLMMs) to study the relationship between time since seroconversion and CD4 count in HIV infected individuals. We implemented a SPMM as well as other techniques for smoothing.

4.1 Data and variables

The Multicenter AIDS Cohort study (MACS) followed 369 HIV infected men aged between 23 and 64 in the USA in the mid 1980s (see, Kaslow et al. [9], and Zeger and Diggle [10] for details). Repeated measurements were obtained for each subject on CD4 cell counts, an important immunological marker which measures the body's ability to fight off infections [24]. Time, measured in years since the date of seroconversion, was known approximately for each subject from several years before seroconversion and up to 6 years after seroconversion. Observations were taken every six months on average for a total of 2,376 measurements (on average 6.5 measures per patient, varying between 1 and 12). Several other measures such as age at study entry, number of cigarette packs smoked per day, number of sexual partners, a measure of depression status (CESD > 0 indicates possible depression), etc. were also collected. This dataset has been analyzed previously (e.g., by Diggle et al. [25]).

4.2 Data analysis

The relationship between time since seroconversion and CD4 count is not immediately apparent from visual inspection of the scatter plot (see Figure 2(a)), apart from a sparsity of subjects with higher CD4 count after 2 years of seroconversion, suggesting that CD4 count decreases with time after seroconversion. With the goal of identifying the progression of mean CD4 counts as a function of time since seroconversion, we used smoothing techniques within a linear mixed effects model (LMM), considering the square root of CD4 count as continuous outcome. We also considered CD4 count as Poisson outcome



(a) Double Hump Shape (fit)

(b) Double Hump Shape (coverage prob.)

(c) Exponential Shape (fit)

(d) Exponential Shape (coverage prob.)

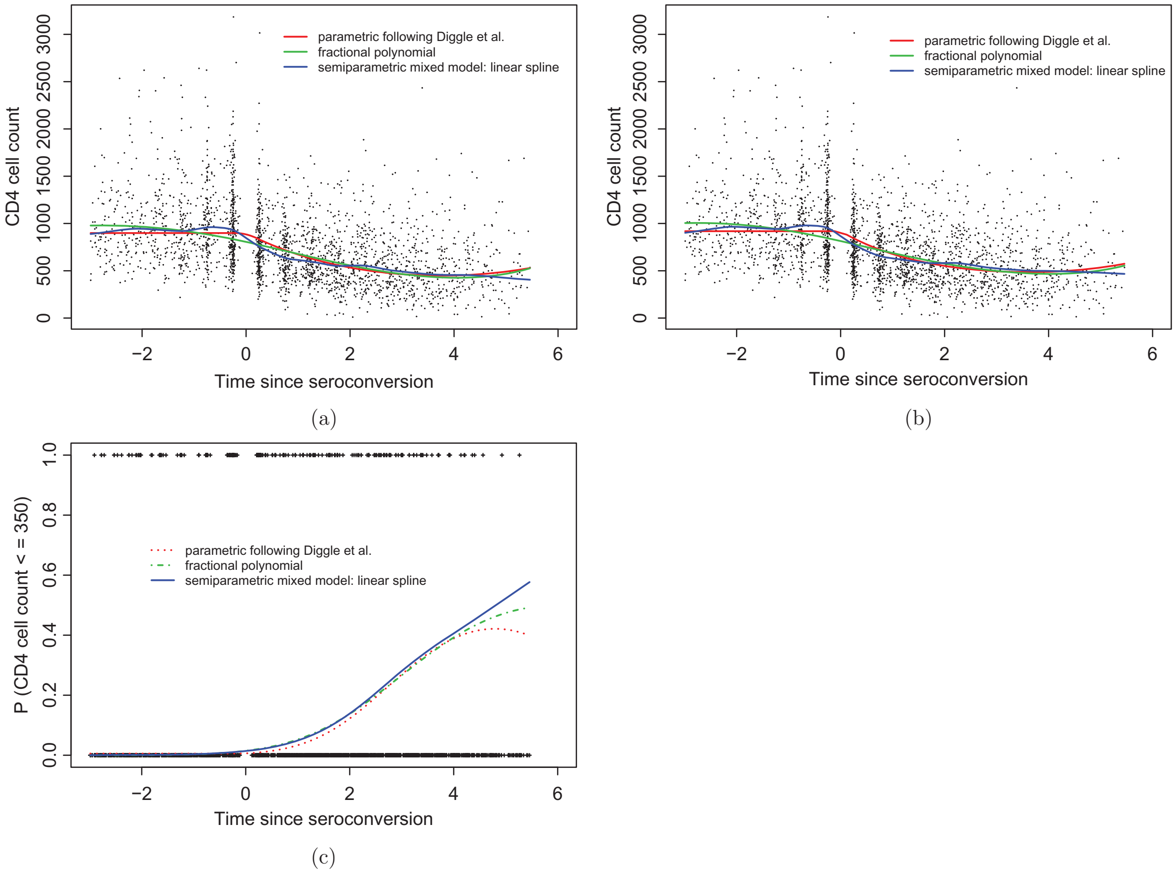
(e) Linear Shape (fit)

(f) Linear Shape (coverage prob.)

Figure 1: True and estimated nonparametric functions based on the mean fitted value at each observed value of x_2 (left) and estimated 95% pointwise coverage probabilities of the true functions (right) for each of the four fits. These results are for normally distributed data with $m = 500$, $n_i = 5$, $\rho = 0.5, \sigma_b^2 = \sigma_\varepsilon^2 = 0.1$ and from 1,000 replications.

and as a binary measure, where $Y = 1$ indicated low CD4 count, i.e., CD4 count ≤ 350 cells/mm^3 and $Y = 0$ indicates CD4 count > 350 cells/mm3. Note that as of March, 2015, the guidelines from the U.S. Department of Health and Human Services suggest starting HIV treatment when one's CD4 count falls to 350 cells/mm^3 or below (www.aids.gov).

To adjust for the effects of potential confounders, smoking (number of packs per day), recreational drug use (yes/no), number of sexual partners, and depression symptoms (as measured by the CESD scale) were also included in all the models as covariates. We centered the number of packs smoked per day, and number of sexual partners at their mean values and included them as linear covariates.



(a)

(b)

(c)

Figure 2: CD4 cell counts with parametric, fractional polynomial, and semiparametric mixed model estimates for the mean trend time obtain from (a) LMM fits, (b) Poisson GLMM fits and (c) Logistic GLMM fits.

Considering the square root of the CD4 count as a continuous response, we constructed the empirical correlation matrix and trajectory plot. Based on these, we adopted a random intercept model with an autoregressive AR(1) correlation structure for the errors. We fit the model

$$g\left(\mathbb{E}\left[Y_{ij}|b_i\right]\right) = f(\text{Time}_{ij}) + \beta_{pack}\ \text{Pack}_{ij} + \beta_{drug}\ \text{Drugs}_{ij} \tag{14}$$

$$+\beta_{partner}\ \text{SexPartner}_{ij} + \beta_{cesd}\ \text{CESD}_{ij} + b_i,$$

where f is some smooth function of time, $i = 1, \ldots, 369$, $j = 1, \ldots, n_i$, $1 \le n_i \le 12$, $\beta_{pack}, \ldots, \beta_{cesd}$ are regression parameters, and $b_i \sim N(0, \sigma_b^2)$.

Assuming the square root of the CD4 cell counts follow a normal distribution, we considered $g\left(\mathbb{E}\left[Y_{ij}|b_i\right]\right) = \mathbb{E}\left[\sqrt{CD4}_{ij}|b_i\right]$ and assumed the random error $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2 \mathbb{V}(\rho))$ such that the correlation between two observations measured at time t_{ij} and t_{ik} for i^{th} subject was $\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}$ for $j \ne k = 1, \ldots n_i$.

Representing CD4 count as Poisson distributed outcome, we fit the GLMM (14) with $g\left(\mathbb{E}\left[Y_{ij}|b_i\right]\right) = \log\left(\mathbb{E}\left[CD4_{ij}|b_i\right]\right)$ and considering AR(1) serial correlation for responses. For binary GLMM, we fit model (14), where $g\left(\mathbb{E}\left[Y_{ij}|b_i\right]\right) = \text{logit}\left(\mathbb{P}\left[CD4_{ij} \le 350|b_i\right]\right)$ and responses were AR(1) correlated. In all cases, we considered the continuous specification of the AR(1) correlation structure that took into account the actual distance between time points.

For each data distribution, we considered four approaches to estimate the nonlinear effect, f of time on CD4 count.

1. A parametric approach, as suggested by Diggle et al. [25]. Following their approach, we specified the mean time-trend for CD4 count as constant prior to seroconversion and quadratic in time thereafter.
2. Using second degree fractional polynomials, where we first identify the best curve within the fractional polynomials framework by ignoring the correlated nature of the data, and then fit the GLMM to respect the correlation structure of the data using the best selected curve.
3. Applying the semiparametric mixed model to smooth the regression spline implementation of the time-CD4 association. We considered linear splines with knots specified as in (13). We also computed the degrees of freedom (df) of the smooth fit following Ruppert et al. [6] (Sections 8.3 and 11.4) to quantify the amount of smoothing used for estimating *f*. A linear function uses 1 df whereas a nonlinear function uses some number greater than 1, depending on the bumpiness of the function.
4. A linear function.

To estimate the model parameters we used the marginal likelihood approach where estimates were obtained by REML for LMM. The parameters of the GLMMs were estimated via PQL method. All analyses were performed in R.

4.3 Results

Table 5 shows the summaries of the baseline characteristics for 369 included men. Mean and standard deviation (SD) are reported for quantitative variables whereas for categorical variables counts and percentages are provided. The mean patient age was 37.4 years.

Table 5: Patients characteristics at baseline (first visit). Mean (SD) is reported for quantitative variables, while count (%) is reported for categorical variables.

Characteristic	Summary Measure ($m = 369$)
Age	37.4 (7.4)
Number of sexual partner	8.9 (3.4)
Smoking (packs of cigarettes/day)	1.2 (1.5)
Recreational drug user	286 (77.5%)
Depressed	181 (49.1%)
CD4 counts	951.7 (400.0)
CD4 counts ≤ 350	271 (11.4%)

Results from all models fits are presented in Table 6. The first panel of Table 6 summarizes results obtained from the mixed effects models fit considering square root of CD4 cell count as a continuous outcome and adopting four different smoothing approaches. The variance and correlation parameter estimates were very similar for all the smoothing methods. The estimated within-subject AR(1) correlation was small (around 0.15). The variance of the random intercept σ_b^2 was estimated as high as around 15 for each of the considered models. The estimated degrees of freedom (df) of the smooth fit was 11.8. The fitted CD4-time trends are displayed in 2 (a) at the mean value of the confounders. The SPMM fit shows that the mean CD4 cell counts remained relatively stable until the time of seroconversion but decreased sharply over the first year after seroconversion. The counts then decreased gradually for the remaining observed time. The FP and parametric (following Diggle et al. [25]) fits did not closely reproduce the behaviour of the SPMM. Indeed, for these two fits there was an increase in CD4 counts after 4 years of the seroconversion. This may be a consequence of the parametric forms of the regression functions or due to a treatment or survivor effect.

The second panel of Table 6 shows results obtained from the models fit considering CD4 cell count as a Poisson outcome. The estimates of the correlation and variance parameters were very similar for all

Table 6: Results from LMMs, Poisson log-normal GLMMs, and logistic-normal GLMMs fits using a specific parametric model, fractional polynomials, semiparametric mixed models, and linear functional form for smoothing.

Smoothing Method	Estimate (95% approximate CI)			
	σ_b^2	σ_ϵ^2	ρ	df
LMM Fit: CD4 as Normal Response				
Parametric	15.23	21.87	0.14	–
(following Diggle et al.)	(12.33, 18.81)	(20.10, 23.80)	(0.10, 0.19)	–
Fractional Polynomial	15.28	21.96	0.13	–
	(12.37, 18.87)	(20.19, 23.88)	(0.10, 0.18)	–
SPMM Approach:	15.35	20.95	0.14	11.80
Linear Spline	(12.48, 18.88)	(19.25, 22.79)	(0.10, 0.19)	
Linear	15.46	22.73	0.15	–
	(12.48, 19.15)	(20.87, 24.76)	(0.11, 0.20)	–
GLMM Fit: CD4 as Poisson Count				
Parametric	0.077	–	0.116	–
(following Diggle et al.)	(0.063, 0.095)	–	(0.082, 0.163)	–
Fractional Polynomial	0.078	–	0.109	–
	(0.063, 0.095)	–	(0.076, 0.155)	–
SPMM Approach:	0.078	–	0.111	11.00
Linear Spline	(0.064, 0.096)	–	(0.077, 0.158)	
Linear	0.078	–	0.120	–
	(0.064, 0.096)	–	(0.085, 0.167)	–
GLMM Fit: CD4 ≤ 350 as 1				
Parametric	4.590	–	0.016	–
(following Diggle et al.)	(3.628, 5.808)	–	(0.007, 0.036)	–
Fractional Polynomial	3.785	–	0.023	–
	(2.930, 4.889)	–	(0.010, 0.052)	–
SPMM Approach:	4.233	–	0.019	3.13
Linear Spline	(3.311, 5.412)	–	(0.008, 0.044)	
Linear	5.127	–	0.013	–
	(4.075,6.450)	–	(0.005, 0.034)	–

σ_b^2 = variability across subjects; σ_ε^2 = random error variance; df = estimated degrees of freedom of the smoother; ρ = serial correlation.

approaches. Fitted curves are shown in Figure 2(b) which are visually indistinguishable from those obtained from models considering CD4 cell counts as continuous.

Results from model fits considering CD4 cell count as a binary response with P(CD4 cell ≤ 350) = 1 are shown in the third panel of Table 6. The estimates of the between-subject variance component σ_b^2 were similar for all the models and quite large on the logit scale for binary outcomes. The estimates of ρ from all the models were near zero indicating no evidence of AR(1) correlation between observations within a subject. The estimated degrees of freedom was 3.13. Figure 2(c) displays the fitted probability of CD4 cell count ≤ 350, considering three smooth curves. For the SPMM fit, it is apparent that the risk of having a CD4 cell count ≤ 350 was zero up until the time of seroconversion, and it increased over time after seroconversion. The models fit using FP and parametric (following Diggle et al. [25]) largely reproduced the trend of SPMM except that they showed a slight decrease in risk after five years of seroconversion.

Simulation results suggested that the correlation and variance parameters were not well-estimated from the linear fit in the presence of curvature. Nevertheless, in our data analysis, the linear fit yielded similar estimates for the correlation and variance parameters to other smoothing approaches. This may be attributed to the fact that the CD4-time association was only slightly nonlinear. However, the shape of the association was better captured by the smoothing methods.

5 Discussion

We examined different approaches to smoothing in generalized linear mixed models and their effects on the estimation of covariance parameters for serially correlated normal, Poisson and binary distributed data through simulation studies. We considered a range of possible exposure-disease association shapes ranging from very simple (linear) to more complex to represent the range of shapes that might be encountered in epidemiologic research. Mixed model representations of penalized splines were applied to estimate smooth functions and compared with other simpler methods for curve fitting such as including a quadratic term, and fractional polynomials.

For normal and Poisson distributed outcomes, we found the semiparametric mixed models (SPMM) performed very well in estimating the correlation and variance parameters as well as the smooth functions even when the true functional form was linear. SPMM outperformed FP and quadratic polynomials, especially when the true association was complex and nonlinear. Our results suggested that proper estimation of the mean structures (fixed effects and smooth functions) was associated with the good estimation of covariance structures (correlation and random effects variances), as expected. The covariance structure of the model is highly dependent on correct specification of the mean structure because the covariance structure only explains the variability not explained by the systematic trend [26].

For binary data, the SPMM exhibited some bias in estimating variance parameters but nevertheless outperformed the other approaches. The bias of the variance component estimates increased as the magnitude of ρ increased. Fitting the true data-generating model also produced biased variance component estimates and were generally no better than the results of fitting the SPMM model, suggesting that the bias was inherent to the estimation method and not attributable to the smoothing approach.

Rather than PQL, using more refined approximation methods, for example Gaussian quadrature, to estimate the GLMMs may reduce the amount of bias in covariance parameters. However, such approximations are not computationally feasible because of the high dimensional integration required to fit the SPMM for generalized responses. Bias-corrected PQL [27] or a Bayesian approach could be more appropriate in binary data situations and is a topic of further research. Nevertheless, PQL is used very frequently because it converges reliably, and can accommodate complex correlation structures (e.g., random effects as well as serial correlations) in most software with reasonable computing time. Chen et al. [8] showed that Gauss Hermite Quadrature performs better than PQL in estimating the variance of random effects in a GLMM for correlated binary data. However, they did not consider the mixed model representation of penalized splines for estimating nonlinear functions. Instead, they used GCV based smoothing splines for estimating. Moreover, they did not consider any serial correlation in the data.

When the model was slightly misspecified (i.e., when data were generated (i) with random intercepts but no serial correlation or (ii) with serial correlation but no random intercepts, but models were fitted that estimated both serial correlation and random intercepts), there was a tendency to underestimate ρ and overestimate σ_b^2 when $\rho > 0$ and $\sigma_b^2 = 0$ and vice versa when $\rho = 0$ and $\sigma_b^2 > 0$ suggesting some uncertainty about capturing the variability in the right form. For normally distributed data, our results were in line with the findings of Demidenko [20]: when the correlation structure was misspecified, there was little difference in the variance parameter estimates provided that the serial correlation was small.

This work has several strengths. Using GLMMs for smoothing in correlated data is currently an active area of research. In this study, a series of simulation studies was carried out across a range of data generation scenarios for three outcome distributions. Complex correlation structures (including both random intercepts and serial correlation) were considered in simulation and data analysis. Further, we compared smoothing via SPMM with other methods in a CD4 cell count dataset considering three different outcome distributions.

This study however has a number of limitations. While a range of scenarios was investigated, the range was not exhaustive and the results from this study may not be applicable to situations not considered. Moreover, we made several simplifying decisions: (i) we smoothed only one covariate; (ii) we generated our covariates to be independent; and, (iii) we used $\sigma_b^2 = 0.25$ for both the Poisson and binary data generation though this value is much bigger on the logit scale for the binary case. Although future studies should address issues related to simultaneously smoothing several covariates, and allowing covariates to be correlated, we believe that understanding the simpler case first is worthwhile. The extension of the SPMM to smooth multiple covariates follows straightforwardly, but may be computationally expensive.

A correctly-specified mean model for response is essential to provide a good estimate of the true dose-response curve. We found the SPMM to be a flexible and useful approach to reveal the nonlinear dose-response relationship in analyzing correlated data. While estimating the nonlinear curve in an elegant way, the SPMM can also estimate the covariance parameters satisfactorily. Moreover, it is a model-based approach to smoothing and incorporation of more complications such as measurement error and missing data is quite straightforward. Although FP and quadratic polynomials performed well in estimating simple nonlinear associations, we recommend using the SPMM as it worked well irrespective of the shape of the association.

Acknowledgements: We thank the reviewers for their thoughtful comments which improved an initial draft of the paper. Andrea Benedetti is supported by the Fonds de recherche du Québec – Santé.

References

1. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 1995;6(4):450–54.
2. Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. Am J Epidemiol 1997;145(8):714–29.
3. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc 1993;88:9–25.
4. Gurrin LC, Scurrah KJ, Hazelton ML. Tutorial in biostatistics: spline smoothing with linear mixed models. Stat Med 2005;24:3361–81.
5. Wand MP. Smoothing and mixed models. Comput Stat 2003;18:223–49.
6. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge: Cambridge University Press, 2003.
7. Lin X, Zhang D. Inference in generalized additive mixed models by using smoothing splines. J Royal Stat Soc: Ser B 1999;61 (2):381–400.
8. Chen J, Liu L, Johnson BA, O'Quigley J. Penalized likelihood estimation for semiparametric mixed models, with application to alcohol treatment research. Stat Med 2013;32:335–46.
9. Kaslow RA, Ostrow DG, Detels R et al. The Multicentre AIDS Cohort Study: rationale, organization and selected characteristics of the participants. Am J Epidemiol 1987;126:310–18.
10. Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to cd4 cell numbers in HIV seroconverters. Biometrics 1994;50: 689–99.
11. Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. J Am Stat Assoc 1998;93:710–19.
12. Rice JA, and Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 2001;57:253–59.
13. Guo W. Functional mixed effects models. Biometrics 2001;58:121–28.
14. Royston P, Altman DG. Regression using fractional polynomials of continuous co-variates: parsimonious parametric modelling (with discussion). Appl Stat 1994;43:429–67.
15. Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of Type I error rate. J Stat Simul Comput 2001;69:89–108.
16. Durban M, Harezlak J, Wand MP, Carrol RJ. Simple fitting of subject-specific curves for longitudinal data. Stat Med 2005;24:1153–67.
17. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Ser B (Stat Methodol) 2011;73(1):3–36.
18. Stone CJ, Koo CY. Additive splines in statistics. Proceedings of the Statistical Computing Section ASA, Washington, DC, 1985;45–48.
19. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Stat Sci 1996;11(2):89–121.
20. Demidenko E. *Mixed Models Theory and Applications*. New Jersey: Willy & Sons, Inc, 2004.
21. Sutradhar BC. *Dynamic Mixed Models for Familial Longitudinal Data*. New York: Springer, 2011.

22. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika 2003;90:455–63.
23. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. New York: Springer, 2000.
24. Alain B. Leucocyte typing: human leucocyte differentiation antigens detected by monoclonal antibodies: specification, classification, nomenclature [report on the first international references workshop sponsored by INSERM, WHO and IUIS]. Berlin: Springer, 1984: 45–48.
25. Diggle PJ, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data*. Oxford: Oxford University Press, 2002.
26. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag, 2000.
27. Lin X. and Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc 1996;91:1007–16.

Supplemental Material: The online version of this article (DOI: 10.1515/ijb-2015-0026) offers supplementary material, available to authorized users. The supplemental files for this article include additional results from the simulation study.