Published in final edited form as:
Biometrics. 2010 March ; 66(1): 79–88. doi:10.1111/j.1541-0420.2009.01240.x.

Variable Selection for Semiparametric Mixed Models in Longitudinal Studies

Xiao Ni[1,*], **Daowen Zhang**[2,**], and **Hao Helen Zhang**[2,***]

[1] Discovery Analytics, GlaxoSmithKline, Five Moore Dr, Research Triangle Park, North Carolina, 27709, USA

[2] Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

Summary

We propose a double-penalized likelihood approach for simultaneous model selection and estimation in semiparametric mixed models for longitudinal data. Two types of penalties are jointly imposed on the ordinary log-likelihood: the roughness penalty on the nonparametric baseline function and a nonconcave shrinkage penalty on linear coefficients to achieve model sparsity. Compared to existing estimation equation based approaches, our procedure provides valid inference for data with missing at random, and will be more efficient if the specified model is correct. Another advantage of the new procedure is its easy computation for both regression components and variance parameters. We show that the double penalized problem can be conveniently reformulated into a linear mixed model framework, so that existing software can be directly used to implement our method. For the purpose of model inference, we derive both frequentist and Bayesian variance estimation for estimated parametric and nonparametric components. Simulation is used to evaluate and compare the performance of our method to the existing ones. We then apply the new method to a real data set from a lactation study.

Keywords

Correlated data; Gaussian stochastic process; Linear mixed models; Smoothly clipped absolute deviation; Smoothing splines

1. Introduction

Semiparametric mixed models (SPMMs; Diggle et al., 2002; Zhang et al., 1998) are useful extension to linear mixed models (Lard and Ware, 1982; Verbeke and Molenberghs, 2000; Diggle et al., 2002) and provide a flexible framework for analysis of longitudinal data. Many authors have studied semiparametric models for longitudinal data in various setup (e.g., Wang, 1998; He et al., 2002; Diggle et al., 2002; Ruppert et al., 2003; Fan and Li, 2004; Chen and Jin, 2006). An SPMM uses parametric fixed effects to represent covariate effects and a smooth function to model the time effect, modeling the within-subject correlation using random effects and stochastic processes. Zhang et al. (1998) adopted a penalized likelihood approach based on smoothing splines, which is computationally efficient and can be conveniently implemented in standard software.

*email: xiao.ni@gsk.com
**email: dzhang2@stat.ncsu.edu
***email: hzhang2@stat.ncsu.edu

In longitudinal studies, often times there are a large number of covariates, but usually not all of them are predictive to the response. For example, in a longitudinal lactation study (Sowers et al., 1993), one hundred fifteen pregnant women were initially enrolled to the study and were scheduled to measure the bone mineral density (BMD) at lumbar spine at four postpartum time points within 18 months. Meanwhile, measurements of many covariates describing participants' physical characteristics, lactation practice and hormonal environment were also taken. One of the objectives of the study is to investigate the pattern of BMD at lumbar spine for postpartum women and to identify from those potential variables the covariates that are associated with the BMD at lumbar spine. Preliminary analysis indicated that on average the BMD at lumbar spine initially declined and then gradually rebounded, and a parametric function may not be adequate to describe this pattern. This motivates the use of an SPMM to model the association of the postpartum BMD at lumbar spine and other covariates and the research to conduct variable selection in SPMMs.

Many variable selection methods have been developed for linear regression models for independent data, such as best subset selection, stepwise selection and shrinkage methods, etc. However, little work had been done for semiparametric models in the longitudinal data settings until the appearance of a recent work of Fan and Li (2004). In the pioneering paper of Fan and Li (2004), they first proposed two estimation procedures to initially estimate regression coefficients: different-based estimator (DBE) and profile least squares. Then they used the local polynomial regression technique to estimate the nonparametric component, and variable selection was achieved by imposing the SCAD (Fan and Li, 2001) penalty on parametric linear covariate effects. As shown in Fan and Li (2004), their methods ignore the correlation in longitudinal data and are therefore very effective in the class of working independent estimators. It is well-known that the estimating equation approach is robust to the mis-specification of the correlation structure in the data when there is no missing data or the missing data mechanism is missing completely at random (MCAR). However, there are some missing data from the motivating lactation study (see Section 6 for more details). To guard against the possible problem associated with missing data, we propose in this paper a double penalized likelihood approach by explicitly taking into account data correlation while conducting model selection and estimation. In particular, our procedure uses random effects to describe the subject-specific effects and uses Gaussian stochastic processes to model the extra temporal correlation among observations within subjects. Since our approach is likelihood based, it will be robust to missing at random (MAR) mechanism and the resulting estimators will be more efficient when the models on the mean and variance structures are correctly specified. Moreover, our method is based on the smoothing spline framework for nonparametric component estimation, which allows us to reformulate the entire problem as a linear mixed effects model (LMM) and hence greatly facilitate the computation. We also show that the variance parameters can be jointly estimated with the smoothing parameters in a unified fashion.

To avoid terminology confusion, we would like to point out that the notion of "double penalty" has been used by various researchers under different contexts. For example, Lin and Zhang (1999) used double penalized quasi-likelihood approach to make inference for generalized additive mixed models, where roughness penalties were imposed on additive nonparametric functions and a quadratic penalty was applied to the random effects. Zou and Hastie (2005) proposed the elastic net for linear model estimation by using a combined penalty of the form $\lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j \beta_j^2$, where β_j are regression coefficients. Lu and Zhang (2006) and Lu (2006) proposed the functional smooth lasso for functional linear model $y_i = \int x_i(t) f(t) dt + \varepsilon_i$ with *i.i.d.* ε_i's, and their procedure has a double penalty on the nonparametric component f as $\lambda_1 \int |f(t)| dt + \lambda_2 \int \{f''(t)\}^2 dt$. Different from all the methods above, our procedure involves a

roughness penalty for the nonparametric component and a shrinkage SCAD penalty for variable selection of parametric components.

The rest of the article is organized as follows. Section 2 first describes the SPMM framework and useful notations. We then introduce the main method, the double-penalized likelihood method for SPMMs. Section 3 describes the computational algorithm for finding the double penalized likelihood estimators. Section 4 derives the variance estimates for parametric and nonparametric components, from both frequentist and Bayesian perspectives. In Section 5 we demonstrate the effectiveness of our method through simulation studies. We illustrate our method through the application to the data from the lactation study in Section 6. We conclude the article with a discussion in Section 7.

2. Double Penalized Likelihood

2.1 Framework and Notation

Suppose that in a longitudinal study there are m subjects, with the ith subject having n_i observations over time. Denote by y_{ij} ($i = 1,\ldots, m, j = 1,\ldots, n_i$) the response at time point t_{ij}. Consider the following semiparametric mixed model

$$y_{ij}=f(t_{ij})+\mathbf{x}_{ij}^T\beta+\mathbf{z}_{ij}^T\mathbf{b}_i+U_i(t_{ij})+\varepsilon_{ij}, \tag{1}$$

where $f(t)$ is an arbitrary twice-differentiable smooth function, β is the $d \times 1$ fixed effects vector of potentially a large number of covariates $\mathbf{x}_{ij}$ from which important covariates are to be selected, $\mathbf{b}_i$ is an $c \times 1$ vector of subject-specific random effects of the corresponding covariates $\mathbf{z}_{ij}$, $U_i(t)$ is an independent mean zero Gaussian process modeling extra underlying serial correlation, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ are independent measurement errors. Further assume $\mathbf{b}_i \sim N\{\mathbf{0}, \mathbf{G}(\varphi)\}$, where $\mathrm{G}(\varphi)$ is a positive definite matrix parametrized by vector φ. The Gaussian process $U_i(t)$ has zero mean and variance-covariance function $\mathrm{cov}\{U_i(t), U_i(s)\} = \gamma(\boldsymbol{\xi}, \alpha; t, s)$ for a specific parametric function $\gamma(\cdot)$ that depends on a parameter vector $\boldsymbol{\xi}$ and α used to characterize the variance and correlation of the process $U_i(t)$. Zhang et al. (1998) considered several forms of Gaussian processes for modeling various within-subject serial correlation. For example, a stationary Ornsterin-Uhlenbeck (OU) process can be used to model homogeneous within-subject covariance structure with constant variance and exponentially decaying correlation: $\mathrm{corr}\{U_i(t), U_i(s)\} = \exp(-\alpha|t - s|)$ (Diggle et al., 2002). If we assume that the variance function changes over time, for instance, in the form of $\exp(\xi_0 + \xi_1 t)$, then the OU process generalizes to a non-homogeneous Ornsterin-Uhlenbeck (NOU) process. We further assume that $\mathbf{b}_i$, $U_i(t)$ and ε_{ij} are mutually independent. Note that an SPMM does not have to contain every single term as given in (1), and it can be easily extended to a more complicated model such as the one where ε_{ij}'s are correlated with a parametric variance-covariance matrix.

We define some matrix notations for convenience. Define $\mathbf{Y}_i = (y_{i1},\ldots, y_{in_i})^T$, and $\mathbf{X}_i$, $\mathbf{Z}_i$, $\mathbf{U}_i$, $\boldsymbol{\varepsilon}_i$ similarly ($i = 1,\ldots, m$). The total number of observations is $n=\sum_{i=1}^m n_i$. Let $\boldsymbol{t}^{\mathbf{0}}=(t_1^0,\ldots, t_r^0)^T$ be an $r \times 1$ vector of ordered distinct values of $\{t_{ij}\}$'s, and $\mathbf{N}$ be the incidence matrix mapping $\boldsymbol{t}^0$ to $\{t_{ij}\}$. Further denote $\boldsymbol{Y}=(Y_1^T,\ldots, Y_m^T)^T$ and $\mathbf{X}$, $\mathbf{N}$, $\boldsymbol{\varepsilon}$ similarly, and $\mathbf{Z} = \mathrm{diag}\{\mathbf{Z}_1,\ldots, \mathbf{Z}_m\}$. Then model (1) can be written in matrix format as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{f} + \mathbf{Z}\mathbf{b} + \mathbf{U} + \boldsymbol{\varepsilon}$, where $\mathbf{f}=\{f(t_1^0),\ldots, f(t_r^0)\}^T$, $\mathbf{b}=(\mathbf{b}_1^T,\ldots, \mathbf{b}_m^T)^T \sim \mathrm{N}\{\mathbf{0}, \mathcal{G}(\varphi)\}$ with variance-covariance matrix $\mathcal{G}(\varphi) = \mathrm{diag}\{\mathbf{G},\ldots, \mathbf{G}\}$; $\mathbf{U}=(\mathbf{U}_1^T,\ldots, \mathbf{U}_m^T)^T \sim \mathrm{N}\{0, \boldsymbol{\Gamma}(\boldsymbol{\xi}, \alpha)\}$ with variance-covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\xi}, \alpha) = \mathrm{diag}\{\boldsymbol{\Gamma}_1,\ldots, \boldsymbol{\Gamma}_m\}$ and the (j, j')th element ($j, j' = 1,\ldots, n_i$) of $\boldsymbol{\Gamma}_i$ being $\gamma(\boldsymbol{\xi}, \alpha; t_{ij}, t_{ij'})$; and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I}_n)$.

2.2 Double Penalized Likelihood

For fixed variance components $\boldsymbol{\theta} = (\boldsymbol{\varphi}^T, \boldsymbol{\xi}^T, \alpha, \sigma^2)^T$, the log-likelihood function of $(\beta, \mathbf{f})$ is (up to a constant)

$$\ell(\beta, \mathbf{f}; \boldsymbol{Y}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\boldsymbol{Y} - \mathbf{X}\beta - \mathbf{N}\mathbf{f})^T \mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\beta - \mathbf{N}\mathbf{f}), \tag{2}$$

where $\mathbf{V} = \text{diag}\{\mathbf{V}_1, \ldots, \mathbf{V}_m\}$ and $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \boldsymbol{\Gamma}_i + \sigma^2 \mathbf{I}_{n_i \times n_i}$. To achieve both model selection and estimation in SPMM (1), we propose to maximize the double penalized (log)likelihood (DPL) function:

$$\ell_{dp}(\beta, f(\cdot); \boldsymbol{Y}) = \ell(\beta, \mathbf{f}; \boldsymbol{Y}) - \frac{\lambda_1}{2}\int_{T_1}^{T_2} \{f''(t)\}^2 dt - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3}$$

where $\lambda_1 > 0$ is a smoothing parameter controlling the balance between the goodness of fit and the roughness of the estimated $f(t)$, and T_1 and T_2 specify the range of t; $p_{\lambda_2}(\cdot)$ is a shrinkage penalty function with $\lambda_2 > 0$ controlling the amount of shrinkage. In order for the use of same λ_2 for all β's to be reasonable, all covariates $\mathbf{x}$ are assumed to be standardized. We call the DPL maximizers $(\hat{\beta}, \hat{f})$ maximum double-penalized likelihood estimators (MDPLEs). It is easy to show that $\hat{f}$ is a natural cubic spline estimate for $f(t)$. There are many choices for the shrinkage penalty $p_{\lambda_2}(\cdot)$ in (3), and we adopt the smoothly clipped absolute deviation penalty (SCAD) due to its desirable theoretical properties (Fan and Li, 2001,2004). The SCAD satisfies $p'_{\lambda_2}(\omega) = \lambda_2[I(\omega \le \lambda_2) + (a\lambda_2 - \omega)_+ / \{(a-1)\lambda_2\} I(\omega > \lambda_2)]$, where $\omega > 0$ and $a > 1$ is another tuning parameter, usually taken a priori (Fan and Li, 2001). Selection of two tuning parameters λ_1 and λ_2 usually needs a two-dimensional grid search, which can be time consuming in practice. In order to reduce the computational cost on parameter tuning, we reformulate the problem using a linear mixed model (LMM) representation; in the new representation, λ_1 is treated as the inverse of a variance component and estimated jointly with other variance components using the REML approach. As shown in Section 3, our procedure only requires one-dimensional parameter tuning on λ_2. We discuss the details in the next section.

3. Computational Algorithm

In this section, we formulate a linear mixed model (LMM) representation for the SPMM, and describe the computational procedures for obtaining parameter estimates based on this LMM representation.

3.1 Linear Mixed Model Representation

By the fact that $\hat{f}(t)$ is a cubic smoothing spline and theorem (2.1) of Green and Silverman (1994), the DPL function (3) can be re-written as

$$\ell_{dp}(\beta, f(\cdot); \boldsymbol{Y}) = \ell(\beta, \mathbf{f}; \boldsymbol{Y}) - \frac{\lambda_1}{2}\mathbf{f}^T \mathbf{K} \mathbf{f} - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{4}$$

where $\mathbf{K}$ is the non-negative definite smoothing matrix. Following Green (1987), we have $\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{B}\mathbf{a}$, where $\mathbf{T} = [\mathbf{1}, \boldsymbol{t}^0]$ and $\mathbf{1}$ is an $r \times 1$ vector of 1's, $\boldsymbol{\delta}$ and $\mathbf{a}$ are 2×1 and $(r-2) \times 1$ vectors

respectively, and $\mathbf{B} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$ with $\mathbf{L}$ being an $r\times(r-2)$ full rank matrix satisfying $\mathbf{K} = \mathbf{L}\mathbf{L}^T$ and $\mathbf{L}^T\mathbf{T} = 0$. Note that $\mathbf{f}^T\mathbf{K}\mathbf{f} = \mathbf{a}^T\mathbf{a}$ and it yields an equivalent double-penalized log-likelihood

$$\ell_{dp}(\beta,\delta,\mathbf{a};\mathbf{Y}) = -(1/2)\log|\mathbf{V}| - (1/2)(\mathbf{Y}-\mathbf{X}_*\beta_*-\mathbf{B}_*\mathbf{a})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}_*\beta_*-\mathbf{B}_*\mathbf{a}) - (\lambda_1/2)\mathbf{a}^T\mathbf{a} - n\sum_{j=1}^{d}p_{\lambda_2}(|\beta_j|)$$

, where $X_* = [\mathbf{NT}, \mathbf{X}]$, $\mathbf{B}_* = \mathbf{NB}$, $\boldsymbol{\beta}_* = (\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$. For fixed $\boldsymbol{\beta}_*$ and given λ_1, λ_2 and $\boldsymbol{\theta}$, maximizing the new DPL with respect to $\mathbf{a}$ gives $\widehat{\mathbf{a}}=(\mathbf{B}_*^T\mathbf{V}^{-1}\mathbf{B}_*+\lambda_1 I)^{-1}\mathbf{B}_*^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}_*\beta)$. Substituting $\hat{\mathbf{a}}$ back to the new DPL, some algebra (Harville, 1977) then leads to an equivalent objective function (up to a constant)

$$\ell_{dp}(\beta_*;\mathbf{Y}) = -\frac{1}{2}(\mathbf{Y}-\mathbf{X}_*\beta_*)^T\mathbf{V}_*^{-1}(\mathbf{Y}-\mathbf{X}_*\beta_*) - n\sum_{j=1}^{d}p_{\lambda_2}(|\beta_j|), \tag{5}$$

where $\mathbf{V}_*=\mathbf{V}+\tau\mathbf{B}_*\mathbf{B}_*^T$. The objective function (5) can be regarded as the penalized log-likelihood function of $\boldsymbol{\beta}_*$ from the following linear mixed model with the same SCAD penalty for $\boldsymbol{\beta}$:

$$\mathbf{Y}=\mathbf{X}_*\beta_*+\mathbf{B}_*\mathbf{a}+\varepsilon_*, \tag{6}$$

where $\boldsymbol{\beta}_* = (\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$ are new fixed effects, $\mathbf{a}$ is treated as random effects distributed as $\mathbf{a} \sim N(0, \tau\mathbf{I})$ with $\tau = 1/\lambda_1$, $\boldsymbol{\varepsilon}_* = \mathbf{Z}\mathbf{b} + \mathbf{U} + \boldsymbol{\varepsilon}$ distributed as $N(\mathbf{0}, \mathbf{V})$, and $\boldsymbol{\theta}_* = (\tau, \boldsymbol{\theta}^T)^T$ are the variance components. It is then equivalent to conducting variable selection for $\mathbf{x}$ in the modified LMM (6). Based on the selected variables and estimated $\hat{\boldsymbol{\beta}}_*$, we can use $\hat{\boldsymbol{\delta}}$ and $\hat{\mathbf{a}}$ to construct the smoothing spline fit $\hat{f}(t)$. The LMM representation suggests that the inverse of the smoothing parameter τ can be treated as a variance component and jointly estimated with $\boldsymbol{\theta}$ using the maximum likelihood or restricted maximum likelihood (REML) approach, which has been discussed by many authors (e.g. Wahba, 1985;Kohn et al., 1991;Speed, 1991;Zhang et al., 1998;Lin and Zhang, 1999).

3.2 Iterative Ridge Regression

For fixed parameters λ_1, λ_2 and $\boldsymbol{\theta}$, direct maximization of (5) is still difficult due to the singularity of the SCAD function. Following Fan and Li (2001,2004), we adopt the iterative ridge regression approach via the local quadratic approximation (LQA). For a small constant ξ and an initial value $|\widehat{\beta}_{*j}^0| \geq \xi$, we have $\{p_{\lambda_2}(|\widehat{\beta}_{*j}|)\}' = p'_{\lambda_2}(|\widehat{\beta}_{*j}|)\mathrm{sgn}(\widehat{\beta}_{*j}) \approx \left\{p'_{\lambda_2}(|\widehat{\beta}_{*j}^0|)/|\widehat{\beta}_{*j}^0|\right\}\widehat{\beta}_{*j}$. Taylor expansion leads to the following optimization problem

$$\max_{\beta}\ell_{dp}(\widehat{\beta}_*|\widehat{\beta}_*^0) \approx -\frac{1}{2}(\mathbf{Y}-\mathbf{X}_*\widehat{\beta}_*)^T\mathbf{V}_*^{-1}(\mathbf{Y}-\mathbf{X}_*\widehat{\beta}_*) - \frac{1}{2}\widehat{\beta}_*^T\sum\nolimits_{\lambda_2}(\widehat{\beta}_*^0)\widehat{\beta}_*,$$

where $\sum_{\lambda_2}(\beta)=\mathrm{diag}\{0, 0, p'_{\lambda_2}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda_2}(|\beta_d|)/|\beta_d|\}$. For fixed $\boldsymbol{\theta}_* = (\tau, \boldsymbol{\theta}^T)^T$ apply the Newton-Raphson method to maximize $\ell_{dp}(\widehat{\beta}_*|\widehat{\beta}_*^0)$ and get the updating formula

$$\widehat{\beta}_*=\left\{\mathbf{X}_*^T\mathbf{V}_*^{-1}\mathbf{X}_*+n\sum\nolimits_{\lambda_2}(\widehat{\beta}_*^0)\right\}^{-1}\mathbf{X}_*^T\mathbf{V}_*^{-1}\boldsymbol{Y}, \tag{7}$$

where $\mathbf{V}_*^{-1}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{B}_*(\tau^{-1}\mathbf{I}_r+\mathbf{B}_*^T\mathbf{V}^{-1}\mathbf{B}_*)^{-1}\mathbf{B}_*^T\mathbf{V}^{-1}$, which is computationally more efficient when r (number of distinct t_i's) is much smaller than n (total sample size). Following each updating step, we set $\hat{\beta}_j=0$ for $|\hat{\beta}_j|<\eta$, where η is a small threshold, say $\eta=10^{-5}$. The non-zero coefficients correspond to important covariates. The iterative ridge regression converges when $\max_j\{|\widehat{\beta}_{*j}^{(k)}-\widehat{\beta}_{*j}^{(k-1)}|/|\widehat{\beta}_{*j}^{(k-1)}|\}<tol$, where tol is a small tolerance.

Remark 1—Very recently, (Zou and Li, 2008) proposed the local linear approximation (LLA) algorithm to solve a SCAD problem. Applying the LLA algorithm to (5) leads to the following optimization problem

$$\max_{\beta}\ell_{dp}(\widehat{\beta}_*|\widehat{\beta}_*^0)\approx-\frac{1}{2}(\boldsymbol{Y}-\mathbf{X}_*\widehat{\beta}_*)^T\mathbf{V}_*^{-1}(\boldsymbol{Y}-\mathbf{X}_*\widehat{\beta}_*)-\sum_{j=1}^{d}p'_{\lambda_2}(|\widehat{\beta}_j^0|)|\beta_j|,$$

which provides an alternative way to solve the problem.

3.3 Iterative Variable Selection Algorithm

We now outline an iterative algorithm that alternatively eliminates unimportant variables and updates parameter estimates. First we fit the full linear mixed model (6) in SAS by including all covariates in the model, and compute the initial values ($\hat{\boldsymbol{\beta}}_{*[1]}$, $\hat{\mathbf{a}}_{[1]}$) and ($\hat{\tau}_{[1]}$, $\hat{\boldsymbol{\theta}}_{[1]}$). For a given λ_2, we propose the following iterative variable selection algorithm:

Step 1—Initialize with $s=1$ and ($\hat{\boldsymbol{\beta}}_{*[1]}$, $\hat{\mathbf{a}}_{[1]}$, $\hat{\tau}_{[1]}$, $\hat{\boldsymbol{\theta}}_{[1]}$).

Step 2—Compute $\hat{\boldsymbol{\beta}}_{*[s+1]}$ using the iterative ridge regression (7) based on current values of $\hat{\boldsymbol{\beta}}_{*[s]}$ and ($\hat{\tau}_{[s]}$, $\hat{\boldsymbol{\theta}}_{[s]}$).

Step 3—Obtain $\hat{\tau}_{[s+1]}$ and $\hat{\boldsymbol{\theta}}_{[s+1]}$ using REML based on important variables.

Step 4—Let $s=s+1$. Go to *Step 2* until the selected covariates converge to a stable set.

Now we describe the REML estimation procedure in *Step 3*. Denote by $\mathbf{X}_{[s]}$ the subset of important variables selected from $\mathbf{X}$ at the sth iteration. The REML log-likelihood of (τ, $\boldsymbol{\theta}$) at this iteration is

$$\ell_R^{[s]}(\tau,\theta;\boldsymbol{Y})=-\frac{1}{2}\log|\mathbf{V}_*|-\frac{1}{2}\log|\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{X}_{*[s]}|-\frac{1}{2}(\boldsymbol{Y}-\mathbf{X}_{*[s]}\widehat{\beta}_{*[s]})^T\mathbf{V}_*^{-1}(\boldsymbol{Y}-\mathbf{X}_{*[s]}\widehat{\beta}_{*[s]}),$$

where $\mathbf{X}_{*[s]}=[\mathbf{NT}, \mathbf{X}_{[s]}]$, and $\widehat{\beta}_{*[s]}=(\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{X}_{*[s]})^{-1}\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\boldsymbol{Y}$ is the MLE of $\boldsymbol{\beta}_*$ based on the selected important variables $\mathbf{X}_{[s]}$. Differentiating $\ell_R^{[s]}(\tau,\mathbf{V};\boldsymbol{Y})$ with respect to τ and θ_k (the kth element of $\boldsymbol{\theta}$) and using the identity $\mathbf{V}_*^{-1}(\boldsymbol{Y}-\mathbf{X}_{*[s]}\widehat{\beta}_{[s]})=\mathbf{V}^{-1}(\boldsymbol{Y}-\mathbf{X}_{[s]}\widehat{\beta}_{[s]}-\mathbf{N}\widehat{\mathbf{f}})$, we obtain the REML estimating equations for τ and $\boldsymbol{\theta}$ (Harville, 1977):

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{B}_*\mathbf{B}_*^T)+\frac{1}{2}(\boldsymbol{Y}-\mathbf{X}_{[s]}\widehat{\beta}_{[s]}-\mathbf{N}\widehat{\mathbf{F}})^T\mathbf{V}^{-1}\mathbf{B}_*\mathbf{B}_*^T\mathbf{V}^{-1}(\boldsymbol{Y}-\mathbf{X}_{[s]}\widehat{\beta}_{[s]}-\mathbf{N}\widehat{\mathbf{F}})=0 \tag{8}$$

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_k})+\frac{1}{2}(\boldsymbol{Y}-\mathbf{X}_{[s]}\widehat{\beta}_{[s]}-\mathbf{N}\widehat{\mathbf{f}})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{V}^{-1}(\boldsymbol{Y}-\mathbf{X}_{[s]}\widehat{\beta}_{[s]}-\mathbf{N}\widehat{\mathbf{f}})=0, \tag{9}$$

where $\mathbf{P}=\mathbf{V}_*^{-1}-\mathbf{V}_*^{-1}\mathbf{X}_{*[s]}(\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{X}_*^T)^{-1}\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}$. Since $\mathbf{V}_*$ is no longer block-diagonal, direct inverse of $\mathbf{V}_*$ may be intractable. We can tackle this problem by using the identity: $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathcal{H}\mathbf{D}^{-1}\mathcal{H}^T\mathbf{V}^{-1}$, where $\mathcal{H}$ is the coefficient matrix of the following system of equations

$$\begin{bmatrix}\mathbf{X}_*^T\mathbf{W}\mathbf{X}_* & \mathbf{X}_*^T\mathbf{W}\mathbf{B}_*\\ \mathbf{B}_*^T\mathbf{W}\mathbf{X}_* & \mathbf{B}_*^T\mathbf{W}\mathbf{B}_*+\lambda_1\mathbf{I}\end{bmatrix}\begin{bmatrix}\beta_*\\ \mathbf{a}\end{bmatrix}=\begin{bmatrix}\mathbf{X}_*^T\mathbf{W}\boldsymbol{Y}\\ \mathbf{B}_*^T\mathbf{W}\boldsymbol{Y}\end{bmatrix}.$$

The Fisher scoring algorithm can then be used to solve (8) and (9) for τ and $\boldsymbol{\theta}$.

The aforementioned algorithm is based on fixed SCAD tuning parameters. Fan and Li (2001) recommended to use $a = 3.7$ as it minimizes the Bayesian risk, and thus we set $a = 3.7$ in our implementation. We propose to tune λ_2 using the Bayesian Information Criterion (BIC) (Schwarz, 1978). For a fixed λ_2, let $\mathbf{X}_1$ and $\boldsymbol{\beta}_1$ be the covariate matrix and coefficients respectively corresponding to the q variables selected by the iterative variable selection algorithm. Fit LMM (6) using the important variables to get $\hat{\mathbf{Y}}=\mathbf{S}\mathbf{Y}$, where $\mathbf{S}$ is a smoother matrix with $q_1 = trace(\mathbf{S})$. Then $\mathrm{BIC}(\lambda_2) = -2\ell_1 + q_1 \log n$, where $\ell_1 = -(n/2)\log(2\pi) - 1/2 \log|\mathbf{V}| - (1/2)(\mathbf{Y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1-\mathbf{N}\hat{\mathbf{f}})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1-\mathbf{N}\hat{\mathbf{f}})$. The generalized cross validation (GCV; Craven and Wahba, 1979) can also be used for tuning λ_2. Wang et al. (2007) compared BIC and GCV for selecting the SCAD tuning parameter and suggested that BIC leads to consistent model selection, whereas GCV tends to have an overfitting effect. Model selection results in our simulation studies also favor BIC, and therefore we chose BIC for tuning λ_2.

4. Frequentist and Bayesian Standard Errors

We derive the frequentist and Bayesian variance formulas for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$. The proposed variance estimates are evaluated via simulations. From frequentists' point of view, $\mathrm{var}(\mathbf{Y}|t,\mathbf{x})=\mathbf{V}$. At convergence, we can write $\hat{\boldsymbol{\beta}}_*=(\hat{\boldsymbol{\delta}}^T,\hat{\boldsymbol{\beta}}^T)^T$ as an approximately linear function of $\boldsymbol{Y}$: $\widehat{\beta}_*=\left\{\mathbf{X}_*^T\mathbf{V}_*^{-1}\mathbf{X}_*+n\sum_{\lambda_2}(\widehat{\beta})\right\}^{-1}\mathbf{X}_*^T\mathbf{V}_*^{-1}\boldsymbol{Y}\equiv\mathbf{Q}\boldsymbol{Y}$. Let $\mathbf{Q}=(\mathbf{Q}_1^T,\mathbf{Q}_2^T)^T$, where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are partitions of $\mathbf{Q}$ with dimensions corresponding to $(\boldsymbol{\delta}^T,\boldsymbol{\beta}^T)^T$, so that $\hat{\boldsymbol{\delta}}=\mathbf{Q}_1\mathbf{Y}$, and $\hat{\boldsymbol{\beta}}=\mathbf{Q}_2\mathbf{Y}$. The estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathrm{var}}_F(\widehat{\beta})=\mathbf{Q}_2\widehat{\mathrm{var}}(\boldsymbol{Y})\mathbf{Q}_2^T=\mathbf{Q}_2\widehat{\mathbf{V}}\mathbf{Q}_2^T. \tag{10}$$

Denote $\hat{\mathbf{a}}(\boldsymbol{\beta}_*)=\mathbf{S}_a\mathbf{Y}$, where $\mathbf{S}_a=\tilde{\mathbf{A}}(\mathbf{I}-\mathbf{X}_*\mathbf{Q})$ and $\tilde{\mathbf{A}}=(\lambda_1\mathbf{I}+\mathbf{B}_*^T\mathbf{V}^{-1}\mathbf{B}_*)^{-1}\mathbf{B}_*^T\mathbf{V}^{-1}$. Therefore $\hat{\mathbf{f}}=\mathbf{T}\hat{\boldsymbol{\delta}}+\mathbf{B}\hat{\mathbf{a}}=(\mathbf{T}\mathbf{Q}_1+\mathbf{B}\mathbf{S}_a)\mathbf{Y}$ and its variance-covariance matrix estimate is

$$\widehat{\mathrm{var}}_F(\hat{\mathbf{f}}|t,\mathbf{x})=\hat{\sigma}^2(\mathbf{TQ}_1+\mathbf{BS}_a)(\mathbf{TQ}_1+\mathbf{BS}_a)^T. \tag{11}$$

From a Bayesian perspective, the double penalized likelihood structure indicates that $f(t)$ has a prior in the form of $\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{Ba}$, with $\mathbf{a} \sim \mathrm{N}(0, \tau\mathbf{I})$ and a flat prior for $\boldsymbol{\delta}$. It follows from the LQA in Section 3.2 that the important coefficients $\boldsymbol{\beta}$ has a prior with log-density kernel equal to $-\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\lambda_2}\boldsymbol{\beta}/2$. The definition of the SCAD penalty implies that some diagonal elements of $\boldsymbol{\Sigma}_{\lambda_2}$ can be zero, corresponding to those $|\beta_j| > a\lambda_2$. Assume after reordering, $\boldsymbol{\Sigma}_{\lambda_2} = \mathrm{diag}\{\mathbf{0}, \boldsymbol{\Sigma}_{22}\}$, where $\boldsymbol{\Sigma}_{22}$ has positive diagonal elements. Then $\boldsymbol{\beta}$ can be partitioned into $(\beta_1^T, \beta_2^T)^T$, where $\beta_1^{d_1\times 1}$ can be regarded as "fixed" effects and $\beta_2^{d_2\times 1}$ as "random" effects with $\beta_2 \sim \mathrm{N}(\mathbf{0}, \sum_{22}^{-1})$. The matrix $\mathbf{X}$ is partitioned into $[\mathbf{X}_1, \mathbf{X}_2]$ accordingly. Now we reformulate the mixed model (6) as: $\mathbf{Y} = \mathbf{NT}\boldsymbol{\delta} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{B}_*\mathbf{a} + \boldsymbol{\varepsilon}_*$, or $\mathbf{Y} = \mathcal{X}_*\boldsymbol{\gamma} + \mathbf{Z}_*\mathbf{b}_* + \boldsymbol{\varepsilon}_*$, where $\mathcal{X}_* = [\mathbf{NT}, \mathbf{X}_1]$, $\gamma=(\delta^T, \beta_1^T)^T$, $\mathbf{Z}_* = [\mathbf{X}_2, \mathbf{B}_*]$, and $\mathbf{b}_*=(\beta_2^T, \mathbf{a}^T)^T$ are the new random effects distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\sum_b=\mathrm{diag}\{\sum_{22}^{-1}, \tau\mathbf{I}_{r-2}\}$, and $\boldsymbol{\varepsilon}_* = \mathbf{Zb} + \mathbf{U} + \boldsymbol{\varepsilon}$ distributed as $\mathrm{N}(\mathbf{0}, \mathbf{V})$. Following Henderson (1975), the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{b}}_*$ under this new mixed model is estimated by

$$\mathrm{var}\begin{pmatrix}\hat{\gamma}\\ \hat{\mathbf{b}}_* - \mathbf{b}_*\end{pmatrix}=\mathbf{C}_*^{-1}=\begin{pmatrix}\mathcal{X}_*^T\mathbf{V}^{-1}\mathcal{X}_* & \mathcal{X}_*^T\mathbf{V}^{-1}\mathbf{Z}_*\\ \mathbf{Z}_*^T\mathbf{V}^{-1}\mathcal{X}_* & \mathbf{Z}_*^T\mathbf{V}^{-1}\mathbf{Z}_*+\sum_b^{-1}\end{pmatrix}^{-1},$$

where $\mathbf{C}_*$ is the corresponding coefficient matrix of Henderson's mixed model equations. From this it is straightforward to construct the Baysian variance-covariance matrices of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$. Denote by $\mathbf{A}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}$ the matrix formed by the block matrices of $\mathbf{C}_*^{-1}$ corresponding to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Similarly denote by $\mathbf{A}_{\boldsymbol{\delta},\mathbf{a}}$ the matrix corresponding to $\boldsymbol{\delta}$ and $\mathbf{a}$. Then the Bayesian variance-covariance matrices for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ are

$$\mathrm{var}_B(\hat{\beta})=\mathbf{A}_{\beta_1,\beta_2}, \tag{12}$$

$$\mathrm{var}_B(\hat{\mathbf{f}})=[\mathbf{T}, \mathbf{B}]\mathbf{A}_{\delta,\mathbf{a}}[\mathbf{T}, \mathbf{B}]^T. \tag{13}$$

These Bayesian variance estimates can be viewed to account for the biases in $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ due to the imposed penalties (Wahba, 1983). We evaluate and compare the empirical performance of frequentist and Bayesian variance estimates in the next section.

5. Simulation Studies

We conduct simulation to evaluate the performance of our DPL procedure and compare this new procedure with the SCAD and LASSO methods in Fan and Li (2004) in both full data and missing data cases. The SCAD and LASSO estimates are computed as Fan and Li's (2004) method implemented in their original Matlab programs. Following Fan and Li (2004), we set the bandwidth h for local polynomial regression equal to $h_0\times$ interquartile range of observed t_{ij}'s. We varied the value of h_0 and found $h_0 = 0.1$ is a reasonable choice in terms of model

errors (defined later) in most cases. Therefore we used $h_0 = 0.1$ for SCAD and LASSO estimates.

We consider three scenarios in the simulation: full data, missing at random (MAR) data and independent data. For the full data scenario, we simulate longitudinal data from the following semiparametric mixed model:

$$y_{ij}=f(t_{ij})+\boldsymbol{x}_i^T\beta+b_{0i}+U_i(t_{ij})+\varepsilon_{ij}, \tag{14}$$

where $i = 1,\ldots, m$ ($m = 40$ or 60), and $j = 1,\ldots, 10$. We consider a staggered entry design: the subjects are equally divided into 5 groups with each group entering at time point 1 through 5; each subject has 10 equally spaced (5 time points in between) measurements. We choose $f(t) = 4\sin(2\pi t/4)$ as the baseline function. By design, there are 50 knots for the smoothing spline fit, and the time t ranges from 0 to 4. The true regression coefficients are $\boldsymbol{\beta} = (.8, .8, 0, 0, .8, 0, 0, 0)^T$, with eight mutually independent covariates $(x_1,\ldots, x_8)$ generated from a standard normal distribution. The random intercept $b_{0i} \sim N(0, v_1)$ represents among-subject variation, and we set $v_1 = 0.36$. We choose a stationary Ornstein-Uhlenbeck (OU) process $U(t)$ with a constant variance σ_u^2 and an exponentially decaying serial correlation: corr $\{U_i(t), U_i(s)\} = \exp(-\alpha|t - s|) = \rho^{|t-s|}$, where $\rho = 0.4$ and $\sigma_u^2=0.5$. We also simulate the measurement error ε_{ij} from $N(0, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon^2=0.25$. In order to investigate the impact on the performance of the DPL under a mis-specified correlation structure, we also apply the proposed DPL approach by assuming a random slope (of t) instead of the random intercept in model (14). This approach is labeled as DPL-MIS in Table 1.

For missing data scenario, we maintain $m = 40$ and simulate MAR data using the main model (14) and the drop-out model in Section 13.3 of Diggle et al. (2002). The ith ($i = 1, 2, \cdots, 40$) subject has no missing value in the first six observations $y_{i1},\ldots, y_{i6}$ and for $j = 7,\ldots, 10$, the drop-out probability p_{ij} of subject i at the jth time point depends on the last previous observation through a logit model: $\text{logit}(p_{ij}) = -1 - 2y_{i,j-1}$. This yielded approximately 30% missing data out of the total 400 observations in the full data. For independent data scenario, we remove the random intercept and the stochastic process in model (14) and simulate independent data for $m = 40$, and then apply the DPL approach assuming an over-parametrized SPMM with a random intercept.

Following Wang et al. (2007), we use the median relative model error (MRME) to evaluate the overall performance of a procedure for model selection and estimation in Table 1. By the simulation design described above, we define the overall model error (ME) of a procedure as

$$\text{ME}=(\widehat{\beta}-\beta)^T(\widehat{\beta}-\beta)+\int_0^T\{\widehat{f}(t)-f(t)\}^2dt/T,$$

where $[0, T = 4]$ specifies the range of the variable t, and the second term is approximated by averaging $\{\hat{f}(t)-f(t)\}^2$ over the design knots. The MRME is then the median of ratios of the ME of a selected model to the commone ME of the estimates from the full model fitted by the REML approach of Zhang et al. (1998). To present a more comprehensive picture, we also use other criteria in Table 1 for performance evaluation. The columns labeled by "Corr." and "Inc." denote the average numbers of correct and incorrect zeros in the coefficient estimates over 100 simulation runs. The column labeled by "Under-fit" corresponds to the proportion of excluding any nonzero coefficients. Similarly, we report the proportion of selecting the exact subset

model in the column "Correct-fit" and the proportion of including all three important variables plus some noise variables in the column "Over-fit".

As can be seen from Table 1, DPL performs very well in terms of all evaluation criteria. The DPL out-performs SCAD and LASSO in terms the overall MRME criterion, even for the cases of mis-specified correlation structures. Although DPL occasionally is slightly more likely than SCAD and LASSO to under-fit the true model by shrinking some important coefficients to zero, it gives a very competitive performance. In the case of data with missing at random mechanism, the performance of SCAD and LASSO deteriorates as expected, particularly with respect to MRME due the poor performance of $\hat{f}(t)$ (Figure 2), while the performance of DPL is not affected by the data missing at random. Although the DPL with a mis-specified correlation structure does not perform as well as the correctly specified DPL, it still gives satisfactory performance and outperforms SCAD and LASSO for a moderate sample size ($m = 60$). In the independent data scenario, although SCAD and LASSO have slightly better performance (as expected), the over-parameterized DPL still delivers a very competitive performance in terms of all evaluation criteria. The estimated variance for random intercept for the over-parameterized DPL is 0.002, indicating that the DPL approach is flexible to produce rather accurate variance component estimates with an over-parameterized correlation structure.

Table 2 summaries the estimated $(\beta_1,\beta_2,\beta_5)$, their relative biases, empirical and model-based standard errors and 95% coverage probabilities. Since SCAD and LASSO give similar numerical results, we only report DPL and SCAD results here. Compared with SCAD (and LASSO), our estimated β_j's have smaller biases in all cases except the independent data scenario. The frequentist and Bayesian standard error formulas derived in Section 4 perform well in most cases: they are close to the empirical estimates and the 95% coverage probability rates are around the nominal level. One observation is that the derived model-based standard errors tend to be under-estimated. We observe that when sample sizes increases ($m = 60$), the differences get smaller. This suggests that the derived standard error formulas may work well when sample size is reasonably large. Although not reported here due to space limit, the estimates from DPL with mis-specified correlation structures are still very good compared to those under correct model specification. Table 3 presents the estimated variance components using REML based on the modified LMM. When the correlation model is correctly specified, the biases become smaller when sample size increases. The variance component estimates are biased under mis-specified correlation structures, as we expected.

The estimated baseline function $\hat{f}(t)$ is also evaluated through visualization. We plot and compare the estimated $f(t)$ and point-wise biases by DPL, LASSO and SCAD, for the full data ($m = 40$) and missing at random data scenarios. For our DPL estimate $\hat{f}$, we also plot the point-wise empirical and model-based frequentist and Bayesian standard errors, and coverage probability of 95% confidence intervals. Figure 1(b) shows that in full data ($m = 40$) case, our approach yields smaller overall biases than the LASSO and SCAD. It can be seen from Figure 1(c) and Figure 1(d) that our standard error formulas for $\hat{\mathbf{f}}$ work well. The Bayesian standard errors are always larger in that they account for the biases due to the imposed penalties.

Figure 2 depicts the results for missing at random data. As shown in Figure 2(a) and Figure 2 (b), SCAD and LASSO produced large biases due to dropped out subjects after the sixth time point. In contrast, our DPL method maintains consistent performance in estimation of $f(t)$ despite 30% missing data. In Figure 2(c), we notice the disparity between the empirical and the estimated standard errors after $t \geq 7$ where the dropout starts occurring, which is due to the smaller sample size caused by the missing data.

6. Application to Longitudinal Lactation Study

In this section, we apply the proposed model selection and estimation procedure for SPMMs to the longitudinal data from the lactation study introduced in Section 1. The detailed description of the study can be found in Sowers et al. (1993). Briefly, the study originally enrolled 115 pregnant women with 0 or 1 parity, who had no intent to breast-feed their incoming babies or intended to breast-feed for at least 6 months. The study participants were then scheduled to have BMD at lumbar spine measured at 4 time points after the birth of their babies. The scheduled time points are 2 weeks, 6 months, 12 months and 18 months postpartum. However, due to various logistic reasons, the actual observation times deviate somewhat from the schedule and several participants missed some scheduled visits. For each woman who made the scheduled visit, the information about her physical activity, dietary calcium intake, lactation practice was obtained. At the same time, blood sample was drawn and assay was conducted to measure various serum hormones including prolactin, parathyroid hormone (PTH) and parathyroid hormone-related peptide (PTHrP), traditionally thought to be related to calcium mobilization for lactating mothers. Since only about 56% of the PTHrP measurements were above the detection limit, PTHrP is dichotomized according to whether or not it was above the limit. One of the study objectives is to identify covariates from this pool of variables that are associated with the postpartum BMD at lumbar spine.

Preliminary analysis indicates that random intercept is sufficient to account for the correlation in the postpartum BMD at lumbar spine and that on average it declined in the first 4 months, then gradually rebounded to a level close to that at week 2 postpartum. In other words, a parametric function may not be appropriate to describe the pattern of postpartum BMD at lumbar spine. This motivates us to consider variable selection in the following SPMM

$$y_{ij}=f(t_{ij})+\mathbf{x}_{ij}^T\beta+b_i+\varepsilon_{ij}, \tag{15}$$

where y_{ij} is the postpartum BMD (g/cm^2) at lumbar spine for woman i measured at time t_{ij}, $f(t)$ is a non-parametric function, $\mathbf{x}_{ij}$ is a 11×1 vector of 11 variables measured at $\boldsymbol{t}_{ij}$: lactation practice (breast-feed or partially breast-feed), physical activity index (mets), body mass index (weight in kilogram/square of height in meters), dietary calcium intake (mg), menstruation status, prolactin (ng/ml), PTH (ng/ml), dichotomized PTHrP as well as the baseline age (year) and infant's birth weight (pound), $b_i \sim \mathrm{N}(0, \sigma_b^2)$ is the woman-specific random effect and $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ is the independent residual error. For numerical stability and ease of interpretation, the continuous covariates are centered by their means. Furthermore, centered dietary calcium intake was divided by 1000, centered BMI, physical activity index and prolactin were divided by 100. After excluding missing data, ninety six women remained in the analysis with 313 total observations. The number of distinct time points is 71.

We apply the DPL procedure to model (15), and the following variables are selected: dietary calcium intake, menstruation status, prolactin, PTH, baseline age and infant's birth weight. This finding is consistent to those in the epidemiology literature. The estimates of the corresponding regression coefficients as well as their estimated standard errors are presented in Table 4. It is seen from this table that after adjusting for time effect all selected variables are positively associated with the postpartum BMD at lumbar spine except the baseline age. The estimated nonparametric function $f(t)$ and its 95% pointwise confidence bands given by the frequentist and Bayesian variance estimation are presented in Figure 3. We can see that the BMD in lumber spline decreases in the first six months, starts increasing from the seventh month and reaches a level comparable to the baseline at the end of the study. Frequentist and Bayesian methods give almost identical standard errors, as shown in Figure 3.

7. Discussion

In this paper we have proposed a new double penalized likelihood (DPL) approach for selecting important parametric fixed effects in semiparametric mixed models (SPMM) for longitudinal data. The DPL is equipped with two penalty terms: the roughness penalty for $f(t)$ and a shrinkage penalty for the fixed covariate effects $\boldsymbol{\beta}$. Maximizing the DPL leads to a parsimonious model and a smoothing spline fit for $f(t)$. We cast the SPMM into a modified linear mixed model framework and proposed an iterative variable selection algorithm for the computation. Within the LMM framework, the inverse of the smoothing parameter is treated as an extra variance component and can be conveniently estimated jointly with other variance components using REML approach. Simulations demonstrated that our method gives very competitive performance in terms of variable selection and parameter estimation, compared with the SCAD and LASSO method in Fan and Li (2004). Under the correct model specification of an SPMM, including the correct specification on the conditional mean structure of the longitudinal data given covariates, random effects and stochastic processes, and the variance structures of the random effects and the stochastic process, our method is more efficient in terms of model selection and parameter estimation than other existing methods which ignore the correlation structure of the data and use working independence correlation matrix. Model diagnostics through exploratory analysis and visualizations are often useful to ensure a properly specified variance structure for the data. Furthermore, when the data set contains missing data subject to missing at random, our DPL estimates still perform consistently. The usage of the proposed method is demonstrated through its application to the data from the longitudinal lactation study.

The proposed method is for selecting important parametric fixed effects in an SPMM. In the future we hope to conduct variance component selection as well. We also plan to extend our DPL methodology to generalized semiparametric mixed models to incorporate other types of endpoints and more likelihood structures. We hope to derive theoretical properties for the estimators. As pointed out by a referee, Zou and Li (2008) recently proposed the local linear approximation (LLA) algorithm and a one-step LLA approach for solving penalized log-likelihood, which potentially can be used in place of LQA to get computationally and statistically more efficient DPL estimators.

Acknowledgments

The research of D. Zhang was supported by National Institute of Health R01 CA85848-08. The research of H. Zhang was supported by in part by National Science Foundation DMS-0645293 and by National Institute of Health R01 CA-085848 The authors thank the editor, the associate editor and two referees for their constructive comments that improved the article.

References

Chen K, Jin ZZ. Partial linear regression models for clustered data. Journal of the American Statistical Association 2006;101:195–204.

Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerishe Mathematik 1979;31:377–403.

Diggle, PJ.; Heagerty, P.; Liang, KY.; Zeger, SL. Analysis of longitudinal data. 2. Oxford University Press; Oxford, U. K: 2002.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 2001;96:1348–1360.

Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. Journal of the American Statistical Association 2004;99:710–723.

Green PJ. Penalized likelihood for general semi-parametric regression models. International Statistical Review 1987;55:245–260.

Green, PJ.; Silverman, BW. Nonparametric regression and generalized linear models. Chapman and Hall; London: 1994.

Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 1977;72:320–340.

He X, Zhu Z, Fung WK. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika 2002;89:579–590.

Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics 1975;31:423–447. [PubMed: 1174616]

Kohn R, Ansley C, Tharm D. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameter. Journal of the American Statistical Association 1991;86:1042–1050.

Lard NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–974. [PubMed: 7168798]

Lin X, Zhang D. Inference in generalized additive mixed models by using smoothing splines. Journal of the Royal Statistical Society, Ser B 1999;61:381–400.

Lu, Y. Technical report, Department of Statistics. University of Wisconsin-Madison; 2006. Contributions to functional data analysis with biological applications.

Lu, Y.; Zhang, C. Technical Report 1126, Department of Statistics. University of Wisconsin-Madison; 2006. Spatially adaptive functional linear regression via functional smooth lasso.

Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric regression. Cambridge University Press; Cambridge, New York: 2003.

Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6:461–464.

Sowers M, Corton G, Shapiro B, Jannausch M, Crutchfield M, Smith M, Randolph J, Hollis B. Changes in bone density with lactation. Journal of the American Medical Association 1993;169:3130–3135. [PubMed: 8505816]

Speed T. Discussion of 'blup is a good thing: the estimation of random effects' by G. K. Robinson. Statistical Sciences 1991;6:15–51.

Verbeke, G.; Molenberghs, G. Linear mixed models for longitudinal data. Springer-Verlag; New York: 2000.

Wahba G. Bayesian 'confidence intervals' for the cross-validated smoothing spline. Journal of the Royal Statistical Society, Ser B 1983;45:133–150.

Wahba G. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. Annals of Statistics 1985;13:1378–1402.

Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 2007;94:553–568. [PubMed: 19343105]

Wang YD. Mixed effects smoothing spline analysis of variance. Journal of the Royal Statistical Society, Ser B 1998;60:159–174.

Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. Journal of the American Statistical Association 1998;93:710–719.

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Ser B 2005;67:301–320.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. (with discussion). Annals of Statistics 2008;36:1509–1533. [PubMed: 19823597]

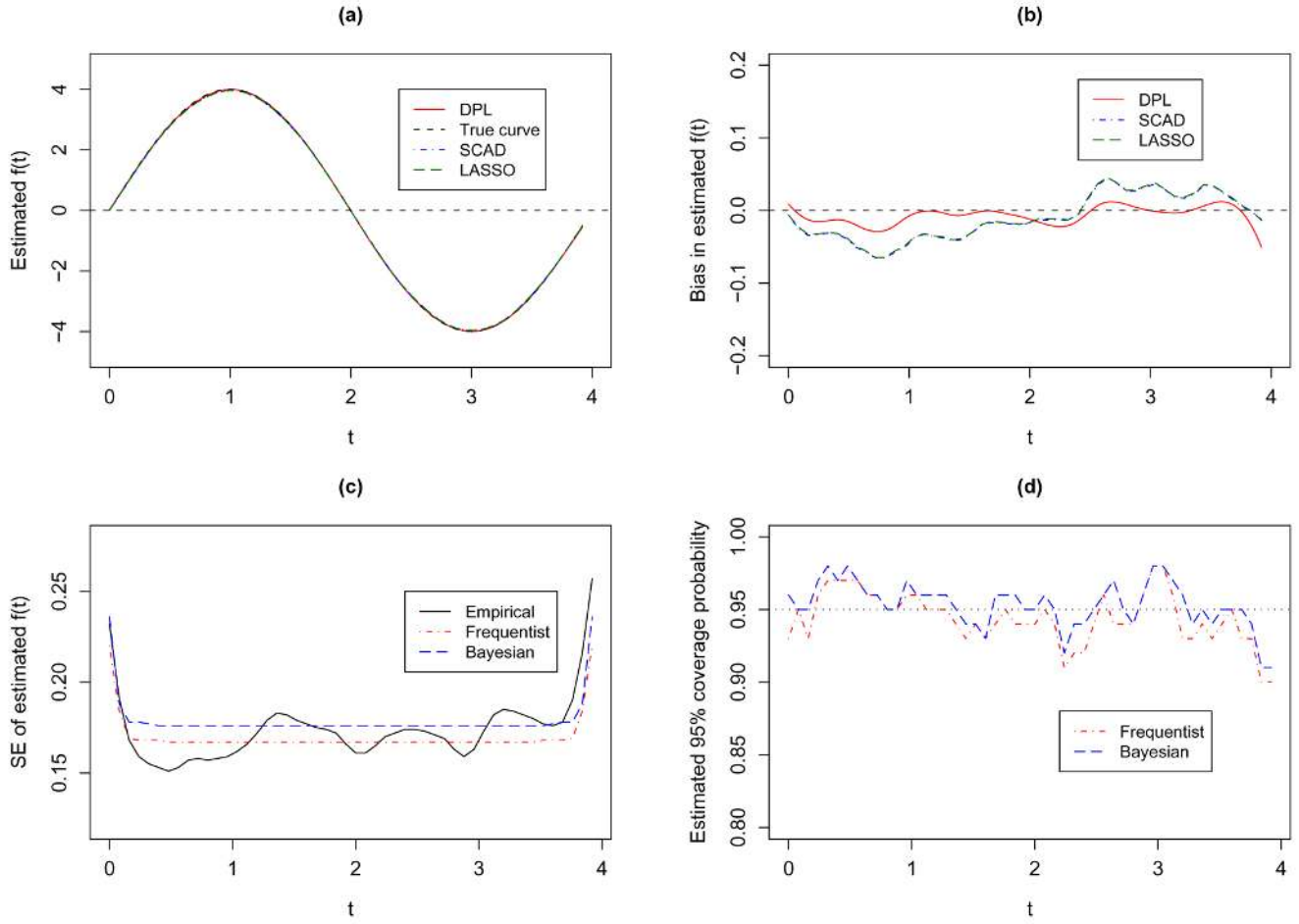

Figure 1.
Plots for estimated $f(t)$ in the full data (m=40) scenario based on 100 samples. Plots (a) and (b) show the averaged fit and point-wise bias by the DPL, SCAD and LASSO methods; plot (c) shows the DPL Bayesian and frequentist standard errors against empirical estimates; and plot (d) plots the averaged coverage probability rates for 95% confidence intervals.

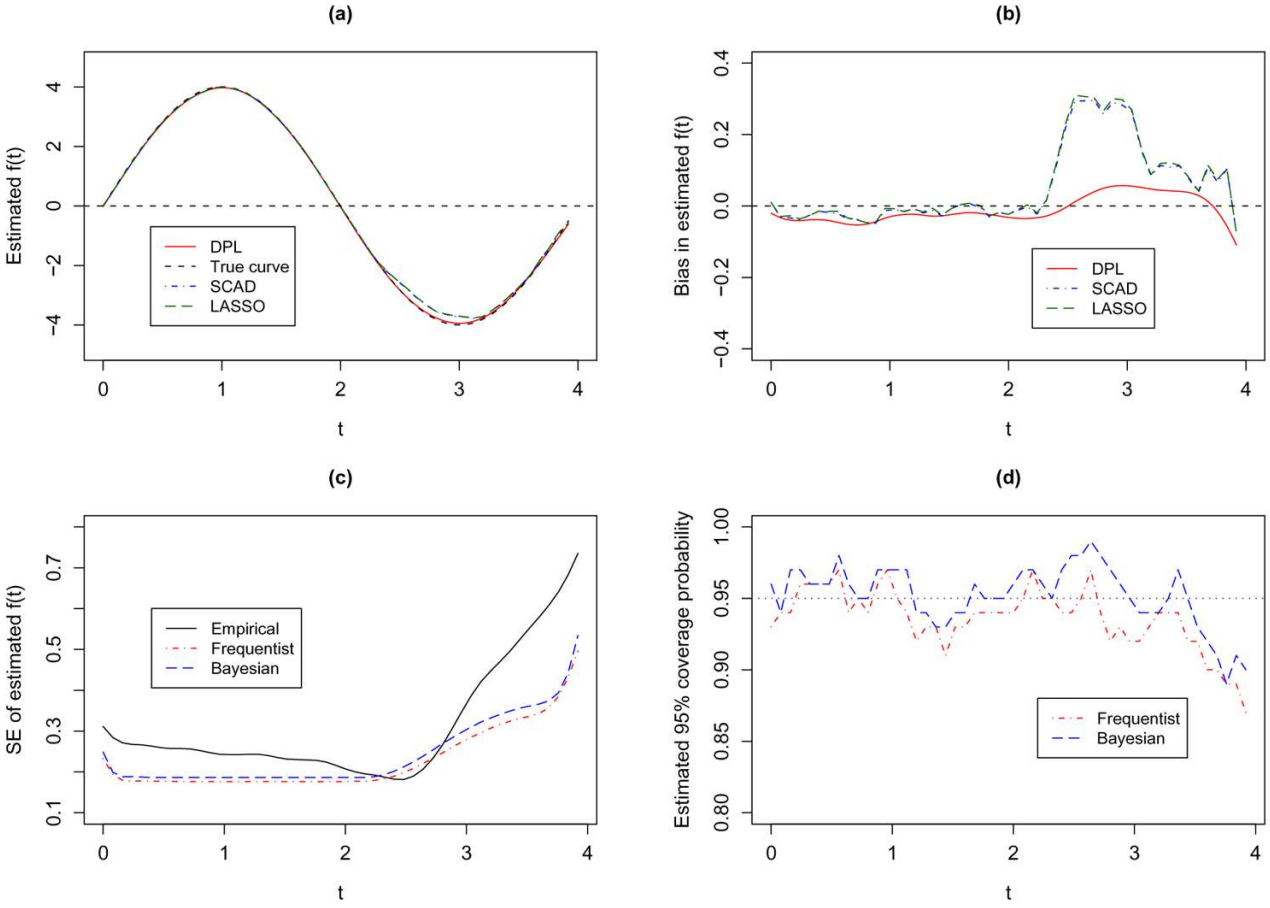

Figure 2.
Plots for estimated $f(t)$ in the missing at random data (m=40) scenario based on 100 samples. Details are as in Figure 1.

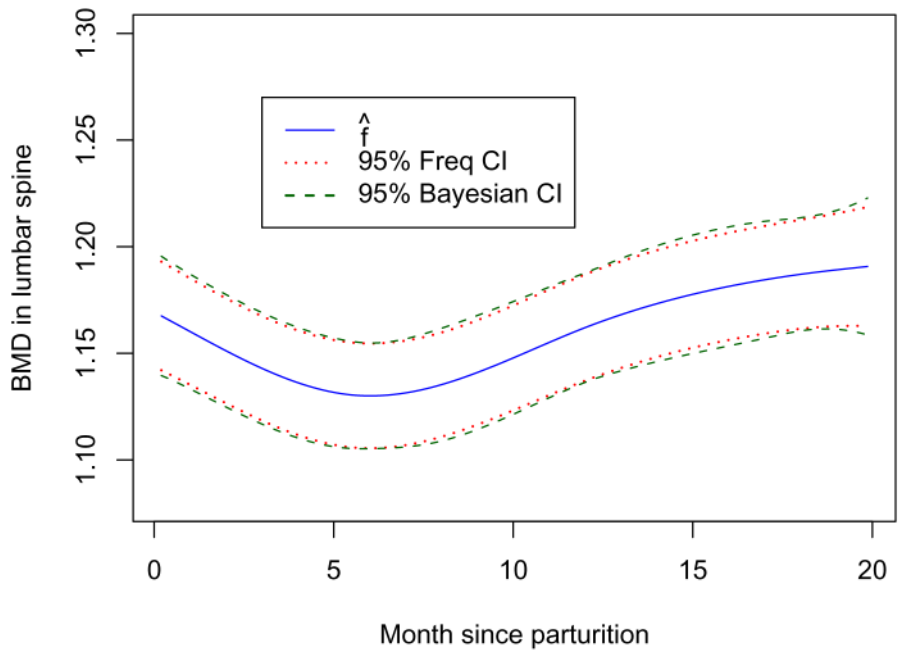

Figure 3.
Plot of estimated baseline functions $f(t)$ in the selected model of (15) and the 95% pointwise confidence intervals. The dotted lines correspond to the frequentist confidence interval, and the dashed lines are given by the Bayesian confidence interval.

Table 1

Model selection and estimation results. model (14). The number in parentheses are standard errors. MRME is the median relative model error in 100 simulation runs. "Corr" and "Inc" respectively denote the average number of correct and incorrect zeros. The number in the parentheses of "Corr. (5)" and "Inc. (0)" are the corresponding measures for the true model.

			Zero coef.		Proportion of		
Scenario	**Method**	**MRME**	**Corr.(5)**	**Inc.(0)**	**Under-fit**	**Correct-fit**	**Over-fit**
Full data $m = 40$	DPL	0.31	4.87	0.04	0.03	0.87	0.10
	DPL-MIS	0.35	4.77	0.03	0.03	0.78	0.19
	SCAD	0.44	4.79	0	0	0.82	0.18
	LASSO	0.43	4.77	0	0	0.81	0.19
Full data $m = 60$	DPL	0.33	4.90	0	0	0.94	0.06
	DPL-MIS	0.32	4.86	0	0	0.91	0.09
	SCAD	0.46	4.76	0	0	0.80	0.20
	LASSO	0.45	4.74	0	0	0.78	0.22
MAR data $m = 40$	DPL	0.24	4.83	0.06	0.04	0.83	0.14
	SCAD	0.58	4.53	0.02	0.02	0.66	0.34
	LASSO	0.59	4.48	0.01	0.01	0.60	0.40
Ind. data $m = 40$	DPL	0.19	4.93	0	0	0.93	0.07
	SCAD	0.14	4.96	0	0	0.96	0.04
	LASSO	0.15	4.96	0	0	0.96	0.04

Table 2

Estimated regression coefficients for important covariates in model (14), standard errors and coverage probabilities (CPs) of 95% confidence intervals

					Model-based SE		95% CP	
Method	**Model Parameter**	**Point Estimate**	**Relative Bias**	**Empirical SE**	**Freq.**	**Bayes.**	**Freq.**	**Bayes.**
Scenario 1: Full data, m = 40								
DPL	β_1	0.789	−0.014	0.167	0.129	0.129	0.93	0.93
	β_2	0.807	0.008	0.156	0.135	0.135	0.92	0.92
	β_5	0.787	−0.016	0.150	0.130	0.130	0.91	0.91
SCAD	β_1	0.775	−0.031	0.141	0.116	-	0.89	-
	β_2	0.789	−0.014	0.141	0.117	-	0.91	-
	β_5	0.764	−0.045	0.145	0.115	-	0.82	-
Scenario 2: Full data, m = 60								
DPL	β_1	0.795	−0.006	0.098	0.103	0.103	0.94	0.94
	β_2	0.807	0.009	0.101	0.105	0.105	0.93	0.93
	β_5	0.795	−0.007	0.105	0.103	0.103	0.93	0.93
SCAD	β_1	0.783	−0.021	0.105	0.096	-	0.90	-
	β_2	0.795	−0.006	0.108	0.097	-	0.90	-
	β_5	0.780	−0.025	0.115	0.097	-	0.89	-
Scenario 3: Missing data, m = 40								
DPL	β_1	0.780	−0.026	0.177	0.144	0.144	0.93	0.93
	β_2	0.795	−0.006	0.151	0.151	0.151	0.96	0.96
	β_5	0.782	−0.022	0.144	0.144	0.144	0.93	0.93
SCAD	β_1	0.765	−0.043	0.163	0.120	-	0.81	-
	β_2	0.776	−0.029	0.169	0.120	-	0.86	-
	β_5	0.755	−0.056	0.169	0.115	-	0.85	-
Scenario 4: Independent data, m = 40								
DPL	β_1	0.773	−0.034	0.045	0.026	0.026	0.71	0.71
	β_2	0.777	−0.029	0.048	0.028	0.028	0.77	0.77
	β_5	0.773	−0.034	0.049	0.027	0.027	0.81	0.81
SCAD	β_1	0.798	−0.002	0.028	0.025	-	0.91	-

Method	Model Parameter	Point Estimate	Relative Bias	Empirical SE	Model-based SE		95% CP	
					Freq.	Bayes.	Freq.	Bayes.
	β_2	0.804	0.005	0.033	0.027	-	0.87	-
	β_5	0.800	0.000	0.028	0.025	-	0.91	-

Table 3

Estimated variance components in the simulation study. Re. bias denotes relative bias.

	Full data (m = 40)		Full data (m = 60)	
Parameter	**Estimate (SD)**	***Re*. bias**	**Estimate (SD)**	***Re*. bias**
DPL				
v_1(intercept)	0.36 (0.22)	0.00	0.33 (0.17)	−0.08
ρ	0.38 (0.23)	−0.05	0.40 (0.20)	0.00
σ_u^2	0.58 (0.19)	0.16	0.56 (0.14)	0.12
σ_ε^2	0.21 (0.11)	−0.16	0.22 (0.09)	−0.12
DPL-MIS				
v_1(slope)	0.08 (0.18)	-	0.06 (0.11)	-
ρ	0.72 (0.09)	0.80	0.72 (0.08)	0.80
σ_u^2	0.85 (0.17)	0.70	0.84 (0.15)	0.68
σ_ε^2	0.38 (0.09)	0.52	0.38 (0.08)	0.52

Table 4

Estimated coefficients and frequentist and Bayesian standard errors under model (15) for data from the longitudinal lactation study

Variable	Full Model $\hat{\beta}(SE_{freq}, SE_{Bayes})$	Selected Model $\hat{\beta}$ (SE_{freq}, SE_{Bayes})
Breastfeed	−0.0203 (0.0091, 0.0092)	0(0, 0)
Partial breastfeed	−0.0020 (0.0090, 0.0091)	0(0, 0)
PTHrP above detection	0.0067 (0.0064, 0.0064)	0(0, 0)
Activity index	−0.0034 (0.0071, 0.0072)	0(0, 0)
BMI	0.1928 (0.1820, 0.1822)	0(0, 0)
Calcium	0.0261 (0.0059, 0.0059)	0.0050 (0.0021, 0.0057)
Menstruation status	0.0244 (0.0076, 0.0076)	0.0101 (0.0040, 0.0072)
Prolactin	0.0173 (0.0057, 0.0057)	0.0055 (0.0021, 0.0053)
PTH	0.0419 (0.0176, 0.0180)	0.0208 (0.0143, 0.0180)
Age	−0.0053 (0.0013, 0.0013)	−0.0013 (0.0001, 0.0013)
Birth weight	0.0404 (0.0046, 0.0046)	0.0305 (0.0028, 0.0039)