Published in final edited form as:

Stat Med. 2013 September 30; 32(22): 3899–3910. doi:10.1002/sim.5790.

A Bayesian Semiparametric Model for Bivariate Sparse Longitudinal Data

Kiranmoy Das¹, Runze Li^{2,3}, Subhajit Sengupta⁴, and Rongling Wu^{3,2}

¹Department of Statistics, Temple University, Philadelphia, PA, 19122

²Department of Statistics, The Pennsylvania State University, University Park, PA, 16802-2111

³Center for Statistical Genetics, The Pennsylvania State University, Hershey, PA 17033

⁴Computer, Information Science & Engineering, University of Florida, Gainesville, FL, 32611

Abstract

Mixed effects models have recently become popular for analyzing sparse longitudinal data that arise naturally in biological, agricultural and bio-medical studies. Traditional approaches assume independent residuals over time and explain the longitudinal dependence by random effects. However, when bivariate or multivariate traits are measured longitudinally, this fundamental assumption is likely to be violated because of inter-trait dependence over time. We provide a more general framework where the dependence of the observations from the same subject over time are not assumed to be explained completely by the random effects of the model. We propose a novel, mixed model based approach and estimate the error-covariance structure nonparametrically under a generalized linear model framework. Penalized splines are used to model the general effect of time and we consider a Dirichlet process mixture of normal prior for the random effects distribution. We analyze blood pressure data from the Framingham Heart Study where body mass index, gender and time are treated as covariates. We compare our method with the traditional methods including parametric modeling of the random effects and independent residual errors over time. Extensive simulation studies are conducted to investigate the practical usefulness of the proposed method. The current approach is very helpful in analyzing bivariate irregular longitudinal traits.

Keywords

Cholesky decomposition; deviance information criterion; Dirichlet process mixture; MCMC; penalized splines

1 Introduction

Most biological experiments result in longitudinal data measured at subject-specific time points. Despite the complexity of modeling the longitudinal measurements, advanced methods of functional and longitudinal data analysis can handle such data effectively. Because measurements from the same subject over different time points are correlated, one needs to model such data carefully. Various parametric, nonparametric and semiparametric methods have been proposed to model the mean trajectory and covariance matrix for irregular sparse longitudinal data (Pan and Mackenzie [1]; Pourahmadi [15, 16]; Das et al. [2]).

Random effects models have been used for analyzing longitudinal traits (Laird and Ware [3]). In the presence of some covariates with fixed effects and others with random effects, mixed effects models have been used in both frequentist and Bayesian approaches to model

longitudinal measurements (Zhang et al. [4]; Zeger and Diggle [5]). Generalized linear mixed models have been proposed to model the response variable in the presence of longitudinal covariates (Ryu, Li and Mallick [6]). Müller et al. [7] developed a semiparametric model to analyze multilevel, repeated measurement data. They used Dirichlet Process Mixture (DPM) priors for subject-specific random effects. Ghosh and Hanson [8] proposed a semi-parametric Bayesian approach for modeling multivariate longitudinal data where they used a mixture of Polya trees prior for subject-specific random effects.

Although all these approaches work reasonably well for their purposes, they all assume that the dependence among the successive measurements from the same subject can be explained completely by the random effects of time in the mixed model. In our opinion, this may be a very strong assumption for multivariate longitudinal data. Different response variables might be internally correlated over time and, hence, it may not be reasonable to assume that the residual errors are independently distributed over time. Also for irregular sparse longitudinal data, it is better not to assume any known parametric structure for the residual covariance matrix.

Flexible distributions have been used to model random effects by many authors in recent literature. Bush and MacEachern [9] and Kleinman and Ibrahim [10] considered a Dirichlet process (DP) prior for the distribution of the random effects. Sethuraman's [11] constructive stick-breaking representation of the DP motivated several authors to consider this specific prior. Recently Li et al. [12] used semiparametric mixed models for longitudinal data by modeling the distribution of random effects with a hierarchically centered DP prior. However, we consider a DP mixture (DPM) of normal prior for the random effects distribution following Ishwaran and James [13].

In this article we propose a mixed model approach for bivariate sparse longitudinal data where residual errors are not assumed to be independent over time. Under a generalized linear model framework, we model the residual covariance matrix nonparametrically. A DPM prior is considered for the distribution of the random effects. Penalized splines (P-splines) are used to model the general effect of time on the traits under consideration. In a Bayesian framework, we derive the full conditional distributions of the model parameters and implement a blocked Gibbs algorithm (Ishwaran and James [13]) for the parameter estimation. We compare our approach with other traditional models in terms of residual sum of squares (RSS) and deviance information criterion (DIC).

We analyze blood pressure (BP) data from the Framingham Heart Study. The study started in 1948 with 5,209 healthy men and women, and its objective was to study the genetic effects on cardiovascular diseases. Currently the study is in its third generation. The data subset analyzed in this paper consists of 977 subjects whose BP, body mass index (BMI), and age are measured longitudinally. We consider BP as the response variable (which is clearly bivariate since we have both systolic and diastolic BP) and other variables are treated as covariates. Figure 1 (and figure 1 in the web appendix) shows the trajectories of systolic (and diastolic) BP for 10 randomly selected subjects. Since there is a correlation between the two response features over time, systolic and diastolic BPs should be modeled jointly for a powerful statistical inference as pointed out by Das et al. [14].

The current article is organized as follows. In Section 2, we describe our model and method for covariance estimation and P-spline for the general effect of time on the bivariate trait. We also discuss the DPM of normal prior for the random effects and the algorithm for parameter estimation. In Section 3, we perform data analysis and compare our results with

the results from other traditional approaches. Section 4 provides the results from the simulation study. Concluding remarks and discussions are given in Section 5.

2 Model and Methods

Consider a random sample of size n drawn from a natural population. We have a bivariate dynamic trait measured longitudinally at subject-specific time points. Let $\mathbf{y}_{ik} = [y_{ik}(t_{i1}), \dots, y_{ik}(t_{iT_i})]$ be the vector of T_i measurements on the i -th subject for the k -th response feature ($k=1$ and 0 say) and $\mathbf{t}_i = (t_{i1}, \dots, t_{iT_i})$ be the vector of measurement-times for the i -th subject. Note that this notation allows subject-specific differences in the number and interval of time points. Assume that there are q covariates correspond to fixed effects ($k=1, 2, \dots, q$) and p ($p < q$) covariates with random effects ($m=1, 2, \dots, p$). The trait value of the response feature k for subject i at time t_{ij} ($j = 1, \dots, T_i$) is modeled as

$$y_{ik}(t_{ij}) = \mathbf{x}_{ijk}^T \boldsymbol{\beta}_k + f_k(t_{ij}) + \mathbf{z}_{ijk}^T \boldsymbol{\gamma}_{ik} + e_{ijk}, \quad (1)$$

where $\mathbf{x}_{ijk}$ is the vector corresponding to the vector of covariates $\mathbf{x}_{ijk}$, and $\mathbf{z}_{ijk}$ is the subject-specific random effects vector associated with the vector of covariates $\mathbf{z}_{ijk}$. Let the smooth function f_k be the general effect of time for the k -th response feature. In matrix notation, model (1) can be represented as

$$Y_{ik} = X_{ik} \boldsymbol{\beta}_k + \mathbf{f}_k(t_i) + Z_{ik} \boldsymbol{\gamma}_{ik} + \mathbf{e}_{ik}, \quad (2)$$

where X_{ik} and Z_{ik} are the corresponding design matrices for the fixed effects and the random effects respectively. The residual vectors $\mathbf{e}_i = [\mathbf{e}_{i1}^T, \mathbf{e}_{i0}^T]^T$ are assumed to be independently normally distributed with mean 0 and covariance matrix $\boldsymbol{\Sigma}_i$. Unlike the traditional approaches, we neither assume that $\boldsymbol{\Sigma}_i$ are identity matrices nor assume any specific form (e.g., AR(1), compound symmetric (CS), CS+AR(1) etc.) for $\boldsymbol{\Sigma}_i$. We will estimate these matrices nonparametrically as described in the following section. We assume that $\mathbf{e}_{ik}$ and $\mathbf{e}_{ik}$ are independent but consider the correlation between $\mathbf{e}_{i1}$ and $\mathbf{e}_{i0}$ (Section 2.3) to explain the dependence between Y_{i1} and Y_{i0} at each time point.

2.1 Regression Model for Covariance Structure

In order to model the covariance matrix, we incorporate a regression based approach (Pourahmadi [15], Pan and Mackenzie [1]). Suppose that a particular subject is measured at T different time points. For the ease of presentation, we suppress the subscript i for the time being. At time t , the residual for the k -th response feature can be modeled in terms of its predecessors in the following way:

$$e_{tk} = \sum_{t'=1}^{t-1} \phi_{t,t'} e_{t'k} + \sum_{t'=1}^{t-1} \psi_{t,t'} e_{t'(1-k)} + \varepsilon_{tk}, \quad (3)$$

where $\phi_{t,t'}$ and $\psi_{t,t'}$ are the corresponding regression coefficients and ε_{tk} is the prediction error with mean 0 and variance $\sigma_{\varepsilon_{tk}}^2$. Note that in model (3), we assume the conditional independence between e_{t1} and e_{t0} given the past times. This model captures the inter-trait and intra-trait correlations over time in an autoregressive way. Denote $\mathbf{e} = [e_1^T, e_0^T]^T$, the vector of residuals of dimension $2T \times 1$, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1^T, \boldsymbol{\varepsilon}_0^T]^T$, where $\boldsymbol{\varepsilon}_k = (\varepsilon_{1k}, \dots, \varepsilon_{Tk})^T$; $k=1, 0$. Then the matrix representation of the above autoregression becomes

$$\mathbf{e} = \begin{bmatrix} \Phi & \Psi \\ \Psi & \Phi \end{bmatrix} \mathbf{e} + \boldsymbol{\varepsilon}, \quad (4)$$

where Φ and Ψ are both $T \times T$ lower triangular matrices with 0s in the diagonal elements and in the (t, t) -th position, $\phi_{t,t}$ and $\psi_{t,t}$ ($t > t$) respectively. We write equation (4) as

$$\mathbf{L}\mathbf{e} = \boldsymbol{\varepsilon}, \quad (5)$$

where $\mathbf{L} = \begin{bmatrix} I - \Phi & -\Psi \\ -\Psi & I - \Phi \end{bmatrix}$. Following Pourahmadi [15], we assume that ε_{it} s are

uncorrelated and hence we get $\text{cov}(\boldsymbol{\varepsilon}) = \mathbf{E} = \begin{bmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_0 \end{bmatrix}$, where both $\mathbf{E}_1$ and $\mathbf{E}_0$ are $T \times T$ diagonal matrices with the t -th diagonal element σ_{t1}^2 and σ_{t0}^2 , respectively. Hence, equation (5) simply gives us

$$\text{cov}(\boldsymbol{\varepsilon}) = \mathbf{L} \text{cov}(\mathbf{e}) \mathbf{L}^T = \mathbf{L} \sum \mathbf{L}^T = \mathbf{E}. \quad (6)$$

Note that from (6), $\Sigma = \mathbf{L}^{-1} \mathbf{E} (\mathbf{L}^{-1})^T$. Thus we have,

$$\begin{aligned} \Sigma^{-1} &= \mathbf{L}^T \mathbf{E}^{-1} \mathbf{L} \\ &= \begin{bmatrix} (I - \Phi)^T & -\Psi^T \\ -\Psi^T & (I - \Phi)^T \end{bmatrix} \begin{bmatrix} \frac{1}{\mathbf{E}^{(1)}} & 0 \\ 0 & \frac{1}{\mathbf{E}^{(0)}} \end{bmatrix} \begin{bmatrix} (I - \Phi) & -\Psi \\ -\Psi & (I - \Phi) \end{bmatrix} \\ &= \begin{bmatrix} \frac{(I - \Phi)^T (I - \Phi)}{\mathbf{E}^{(1)}} + \frac{\Psi^T \Psi}{\mathbf{E}^{(0)}} & -\frac{(I - \Phi)^T \Psi}{\mathbf{E}^{(1)}} - \frac{\Psi^T (I - \Phi)}{\mathbf{E}^{(0)}} \\ -\frac{\Psi^T (I - \Phi)}{\mathbf{E}^{(1)}} - \frac{(I - \Phi)^T \Psi}{\mathbf{E}^{(0)}} & \frac{\Psi^T \Psi}{\mathbf{E}^{(1)}} + \frac{(I - \Phi)^T (I - \Phi)}{\mathbf{E}^{(0)}} \end{bmatrix} \\ &\triangleq \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \end{aligned}$$

In a sparse irregular longitudinal setting, the subject-specific covariance matrix will be

modeled using $\mathbf{E}_i = \mathbf{L}_i \mathbf{\Lambda}_i \mathbf{L}_i^T$. Hence, for i -th subject, we have $\Sigma_i^{-1} = \begin{bmatrix} \Sigma_{11}^{(i)} & \Sigma_{12}^{(i)} \\ \Sigma_{12}^{(i)T} & \Sigma_{22}^{(i)} \end{bmatrix}$.

Following Pourahmadi [15,16], Pan and Mackenzie [1], Daniels and Pourahmadi [17], we express the unconstrained dependence parameters $\log \sigma_{t1}^2$, $\log \sigma_{t0}^2$, $\phi_{t,t}$ and $\psi_{t,t}$ as polynomial functions of time as follows:

$$\log \sigma_{t1}^2 = \lambda_0 + \lambda_1 t + \lambda_2 t^2 + \dots + \lambda_g t^g, \quad (7)$$

$$\log \sigma_{t0}^2 = \delta_0 + \delta_1 t + \delta_2 t^2 + \dots + \delta_g t^g, \quad (8)$$

$$\phi_{t,t'} = \eta_0 + \eta_1 (t - t') + \eta_2 (t - t')^2 + \dots + \eta_h (t - t')^h, \quad (t' = 1, 2, \dots, t-1), \quad (9)$$

$$\psi_{t,t'} = \theta_0 + \theta_1 (t - t') + \theta_2 (t - t')^2 + \dots + \theta_h (t - t')^h, \quad (t' = 1, 2, \dots, t-1). \quad (10)$$

The suitable orders of the above polynomials $\mathbf{g}$ and $\mathbf{h}$ are determined from the data. Pourahmadi [15,16], Pan and Mackenzie [1] proposed likelihood based approaches to estimate the parameters in the covariance matrix using the iterative reweighted least squares (IRLS) algorithm. We will estimate those parameters in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods that are discussed in the later sections. Note that our approach of modeling covariance matrices, which hereafter we denote by “the proposed approach”, is more flexible than the traditional autoregressive (AR(1)) or compound symmetry (CS) structures simply because real data might not display those specific forms.

2.2 Penalized Spline for the General Effect of Time

To model f_k , the general effect of time for the k -th response feature, one can use different smoothing techniques. Typically, smoothing splines method is preferred (Li et al. [12]) simply because it has the smallest mean squared error (MSE). Also, one can avoid the knot selection problem to some extent by using smoothing splines. However, since we have a reasonably large sample, we prefer to use penalized splines (Ruppert, Wand and Carroll [18]) to avoid the computational complexities of smoothing splines for higher dimensions. With a polynomial spline of degree r with knots $(t_1, \dots, t_K)$, response-specific means at different time points as described in equation (1) can be expressed as the following non-parametric regression model:

$$f_k(t_{ij}) = b_{k0} + b_{k1}t_{ij} + b_{k2}t_{ij}^2 + \dots + b_{kr}t_{ij}^r + \sum_{s=1}^K c_{ks}(t_{ij} - T_s)_+^r, \quad (11)$$

where $(x)_+^r = x^r I(x > 0)$ and $(t_1 < t_2 < \dots < t_K)$ is a fixed set of knots. In practice, it is desirable to keep the degree of the spline (r) relatively small, between 1 and 3. Using matrix notation, the above equation can be written as

$$f_{ik} = U_i b_k + V_i c_k, \quad (12)$$

where

$$U_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & \dots & t_{i1}^r \\ 1 & t_{i2} & t_{i2}^2 & \dots & t_{i2}^r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{ir} & t_{ir}^2 & \dots & t_{ir}^r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{iT_i} & t_{iT_i}^2 & \dots & t_{iT_i}^r \end{bmatrix}$$

and

$$V_i = \begin{bmatrix} (t_{i1} - T_1)_+^r & (t_{i1} - T_2)_+^r & \dots & (t_{i1} - T_K)_+^r \\ (t_{i2} - T_1)_+^r & (t_{i2} - T_2)_+^r & \dots & (t_{i2} - T_K)_+^r \\ \vdots & \vdots & \ddots & \vdots \\ (t_{ir} - T_1)_+^r & (t_{ir} - T_2)_+^r & \dots & (t_{ir} - T_K)_+^r \\ \vdots & \vdots & \ddots & \vdots \\ (t_{iT_i} - T_1)_+^r & (t_{iT_i} - T_2)_+^r & \dots & (t_{iT_i} - T_K)_+^r \end{bmatrix}.$$

The coefficients of the spline part represent the measure of roughness at the respective knots and we make the curve smoother by placing a penalty on those parameters. In other words, we would like to minimize the following expression

$$S = \sum_{i=1}^n \sum_{j=1}^{T_i} \sum_{k=0}^1 (y_{ik}(t_{ij}) - f_{ik}(t_{ij}))^2 + \lambda^* \sum_{s=1}^K \sum_{k=0}^1 c_{ks}^2. \quad (13)$$

* is the smoothing parameter having significant control over the smoothing process. In a Bayesian framework, this can be achieved by considering a normal prior for c_k followed by a gamma prior for *. Knots are selected from the evenly spaced sample quantiles for the independent variable (time, in our case) following the traditional approach (Ruppert, Wand and Carroll [18]).

2.3 Priors for Random Effects

DPM random effects models have become quite popular in recent Bayesian literature mainly because (i) they are computationally feasible and (ii) they provide better parameter estimates (with much shorter credible intervals). Since the seminal works by Blackwell and Macqueen [19] and Ferguson [20], DP is being used in various applied statistics literature. Instead of modeling the vector of random effects $\gamma_i = [\gamma_{i1}^T, \gamma_{i0}^T]^T$ as normal variables with unknown but fixed parameters, DPM assumes that the parameters of the normal distribution actually come from a mixture (possibly infinite) density, which we model with a DP.

We follow the approach given in Ishwaran and James [13]. Note that our approach of modelling y_i does not consider the correlation between the two response features Y_{i1} and Y_{i0} at each time point. This correlation is captured by assuming a multivariate normal distribution for y_i . Note that the vectors y_{i1} and y_{i0} are of dimension $(p+1) \times 1$ assuming the design matrix Z_{ik} has all 1's in the first column. Let ρ be the correlation between y_{is1} and y_{is0} , for $s = 0, \dots, p$ and $\text{var}(\gamma_{is1}) = \sigma_{is1}^2$, $\text{var}(\gamma_{is0}) = \sigma_{is0}^2$. Hence our DPM model for the random effects can be represented as:

$$\begin{aligned} \gamma_i &\sim MV N(\text{mean}=\mathbf{0}, \text{Cov}=\Sigma'(\rho, \sigma_{is1}^2, \sigma_{is0}^2; s=0, \dots, p)), \\ (\rho, \sigma_{is1}^2, \sigma_{is0}^2; s=0, \dots, p) &| G \stackrel{\text{iid}}{\sim} G, \\ G &\sim DP(\alpha, G_0). \end{aligned}$$

The concentration parameter α is assumed to be unknown and G_0 is the base distribution of the DP. We note that by assuming a DP prior for the covariance parameters $(\rho, \sigma_{is1}^2, \sigma_{is0}^2; s=0, \dots, p)$, we essentially classify them into several clusters. Let $C_1, C_2, \dots, C_n$ be the classification variables induced by the DP prior for n subjects under study and c denotes a generic value of C_i . Our approach is based on the following hierarchical model:

$$\begin{aligned} \gamma_{isk} | \sigma_{isk}^2, C_i &\sim N(0, \sigma_{isk}^2), \quad \text{for } k=1, 0; \quad s=0, \dots, p, \\ (C_i | p) &\stackrel{\text{iid}}{\sim} \sum_{j=1}^N p_j \delta_j(\cdot), \\ (\sigma_{isk}^2 | \nu_1, \nu_2) &\stackrel{\text{iid}}{\sim} IG(\nu_1, \nu_2), \quad \text{for } k=1, 0; \quad s=0, \dots, p, \\ \rho &\sim \text{Uniform}(-1, 1), \end{aligned}$$

where $IG(\nu_1, \nu_2)$ stands for inverse gamma distribution with mean $\frac{\nu_2}{\nu_1 - 1}$ and the weights p_j are chosen by Sethuraman's [11] stick breaking representation; $p_1 = V_1$, $p_2 = (1 - V_1) V_2$,

....., $p_N = (1 - V_1)(1 - V_2) \dots (1 - V_{N-1})V_N$, where V_i 's ($i=1, 2, \dots, N-1$) are iid Beta(1, α) and $V_N = 1$. We consider $(\alpha|\eta_1^*, \eta_2^*) \sim \text{Gamma}(\eta_1^*, \eta_2^*)$. It can be shown that the distribution for $C_i|p$ converges almost surely to a DP with concentration parameter α . The truncation of DPM to a finite mixture is done by making the expected approximation error arbitrarily small (Ishwaran and James [13]). The full conditional distributions and detailed estimation procedure for blocked Gibbs sampling can be found in Ishwaran and James [13].

2.4 Prior Selection and Posterior Estimation

Let $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\tau}$ denote the vectors of coefficients for the variance structures and dependence structures for two responses, respectively. Also, $\mathbf{b}_k$ and $\mathbf{c}_k$ are the coefficients in f_k , as described earlier ($k=1, 0$). We consider a multivariate normal prior with mean 0 and covariance matrix $\boldsymbol{\Sigma}_b$ for $\mathbf{b}_k$. Note that we take exactly the same prior for $k=0$ and 1. For $\mathbf{c}_k$, we take a multivariate normal prior with mean 0 and covariance matrix $\boldsymbol{\Sigma}_c$, where $\boldsymbol{\Sigma}_c$ is a diagonal matrix with diagonal element $1/\sigma_k^2$. A gamma prior with parameters (α, β) is taken for σ_k^2 . This will tend to shrink the roughness of the spline towards zero which is indeed the Bayesian equivalence of the frequentist approach to penalized spline. Priors for $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\tau}$ are taken as $\text{MVN}(0, \boldsymbol{\Sigma}_\mu)$, $\text{MVN}(0, \boldsymbol{\Sigma}_\sigma)$, $\text{MVN}(0, \boldsymbol{\Sigma}_\tau)$ and $\text{MVN}(0, \boldsymbol{\Sigma}_\tau)$ respectively. A normal prior with mean 0 and covariance matrix $\boldsymbol{\Sigma}_\tau$ is considered for the parameter vector $\boldsymbol{\tau}$.

Priors for the random effects distribution are already specified in Section 2.3. Full conditional distributions for the parameters under the above specification of priors are given in the web appendix. We use a hybrid scheme of Gibbs sampling and Metropolis Hastings (MH) algorithm for the estimation purpose along with blocked Gibbs sampling (Ishwaran and James [13]). For the MH algorithm, the proposal densities are suitably chosen so that the acceptance rate is quite satisfactory (0.20 or above). Sensitivity analysis is performed to investigate the effects of the prior distributions on the estimated parameter values. Convergence of the chains is assessed by computing Multivariate Potential Scale Reduction Factor (Brooks and Gelman [22]).

3 Data Analysis

We analyze BP data from the Framingham Heart Study. As mentioned earlier, the objective of this study was to investigate the effects of genetic factors on cardiovascular diseases. The sample consists of 977 subjects (500 males, and age ranges from 29 to 104), whose BP (systolic and diastolic) and BMI are measured over different time points. Number of measurements are different for different subjects and that lead to sparse irregular longitudinal data. The number of measurements from a specific subject may be as low as 3. The BPs are set to be the response variable and BMI, gender and time to be the covariates. Note that there is an inter-trait correlation (between the systolic and diastolic BP) and hence two response features should be modeled jointly (Das et al. [14]).

Model and Priors

By denoting $k=1$ for systolic BP and 0 for diastolic, we fit the following bivariate mixed model:

$$y_{ijk} = \beta_{0k} + \beta_{1k} x_{ij1} + \beta_{2k} x_{ij2} + f_k(t_{ij}) + \gamma_{i0k} + \gamma_{i1k} t_{ij} + e_{ijk}, \quad (14)$$

where $i=1, \dots, 977$; $j=1, \dots, T_i$; $k=1$ and 0. Note that in the above model, we consider $x_{ij1} = x_{ij01} = \text{BMI}$ and $x_{ij2} = x_{ij02} = \text{gender}$, which we code 1 for male and 0 for female. Hence, β_{0k} denotes the fixed intercept, β_{1k} and β_{2k} denote the effects of BMI and gender on k -th response feature respectively. The general effect of time on the k -th response is modeled by

f_k and the subject-specific variation is captured by the random intercept β_{0k} and the random slope β_{1k} . The random effect of time on the response is considered to be linear since it provides the smallest DIC value compared to quadratic and cubic random effects model (S.1 in the Web appendix).

We consider the following prior distributions for the model parameters. We assume that Σ_b is diagonal with all diagonal elements=1,000, Σ_c is diagonal with all diagonal elements=100. Σ_γ has gamma($\alpha_\gamma = 0.001$, $\beta_\gamma = 1,000$), a diffuse prior distribution (so to speak) and Σ_δ , Σ_ϵ and Σ_η are diagonal matrices with diagonal elements 30, 25, 20 and 10 respectively. For Σ_θ , we consider a diagonal matrix with diagonal elements 10. For the truncated DPM approach, we consider $\eta_1 = \eta_2 = 2$, $\eta_1^* = 1.5$, $\eta_2^* = 2.5$.

Computational Details

Initial values of the parameters are drawn from the respective prior distributions. We run 120,000 iterations and initial 10,000 burn-in iterations are discarded in order to remove the effect of the initial values. The convergence of the chains are assessed by computing multivariate potential scale reduction factors (MPSRF) as proposed by Brooks and Gelman [22]. We consider 5 different chains with different starting points and discard the burn-in iterations. MPSRF checks whether the chains come from the same distribution, which happens if convergence is achieved. Model parameters are estimated by their respective posterior means.

Model Selection

To obtain a reasonable number of knots as suggested by the data, we perform exploratory data analysis. No significant improvement in terms of smoothness of the mean curve is found for more than 5 knots. So, $K=5$ is taken as the required number of knots. To get the optimal triplet (r, g, h) , we compute DIC values for different triplets (r, g, h) . Note that we compute DIC on the basis of the conditional likelihood (conditioned over the random effects) following Hadfield's [23] recommendation. From Table 1, we note that (2,2,2) gives the smallest DIC value and hence considered as the optimal triplet. We take $N=50$ as the truncation value for DPM which gives an acceptable expected approximation error (smaller than 0.01).

Results

Traditionally for analyzing bivariate longitudinal data with mixed models, researchers use (i) independent errors, (ii) $UN \sim AR(1)$ structure or (iii) $UN \sim CS$ structure (Sithole and Jones [24], Das et al. [14]). Note that UN stands for unstructured covariance matrix; in our

case this can be expressed as $UN = \begin{bmatrix} \sigma_s^2 & \sigma_{sd} \\ \sigma_{sd} & \sigma_d^2 \end{bmatrix}$, where σ_s and σ_d denote respectively the standard deviation for systolic and diastolic BP and σ_{sd} denotes the covariance between them. We compare the proposed method with the aforementioned traditional approaches and also with $CS + AR(1)$ method in terms of RSS and DIC. As shown in Table 2, the proposed approach of covariance estimation gives the smallest RSS and DIC when DPM of normal prior is used for the random effects. This table also shows that, in general, normal random effects models provide larger values for RSS and DIC compared to DPM random effects models. Figure 2 (and figure 2 in the web appendix) shows the fitted curves with the proposed covariance estimation and the traditional approach of independent residuals model. Random effects of time are modeled by DPM of normal distribution in both cases. Solid curves are the resulting estimates of the proposed method, dotted curves are from the model with independent residuals over time (traditional approach). Proposed approach provides a better fit. Figure 3 (and figure 3 in the web appendix) shows the estimated mean curves for

DPM random effects model (solid curves) and normal random effects model (dotted curves). Proposed covariance estimation method is used in both cases. Here also DPM random effects model gives a better fit compared to normal random effects model.

Figure 4 shows the fitted curves for the systolic and diastolic BP using the proposed approach. Note that while the systolic BP shows a slightly increasing trend over time, the diastolic BP is more or less stable with a very small change (from approx. 80 to 90). This shows that the diastolic blood pressure for Framingham cohort is controlled over time and systolic component is increasing but below 150 till age 80. In general this cohort seems to have controlled BP and hence is not at risk for cardiovascular diseases due to uncontrolled blood pressure. The parameter ρ shows the correlation between the two response features over time. For the current data, ρ is estimated as 0.57 with (0.23, 0.82) a 95% credible interval, a strong correlation between two response features at each time point. Figure 5 shows the estimated posterior density for ρ .

Das et al. [14] analyze this dataset to locate significant SNPs controlling human BP over time. They discuss the advantage of joint modeling of systolic and diastolic BP over the separate modeling. They compare joint and separate analysis in terms of power and false positive rate (FPR) by extensive simulation studies and show that for such studies, joint analysis gives more power and less FPR and hence provides more reliable inference. Following their recommendation, we perform a joint analysis for these two correlated response features in the current article.

4 Simulation Study

For comparing the performance of the DPM random effects model to the traditional normal random effects model, we perform simulation studies. We simulate bivariate response data using equation (14) with the estimated model parameters from the Framingham Heart Study (FHS). Assume the residual vectors $e_j \sim MVN(0, \Sigma_j)$, where Σ_j are modeled with equation (6) with the estimated parameter values for ρ , σ_{00}^2 , σ_{01}^2 and σ_{10}^2 from the analysis of the FHS data.

We consider a mixture normal distribution for the vector of random effects γ_j . For $s = 0, 1$ and $k = 0, 1$; let ρ be the correlation between γ_{is1} and γ_{is0} , $var(\gamma_{isk}) = \sigma_{sk}^2$. Vectors of random effects are simulated from the mixture density $f = 0.5 f_1 + 0.3 f_2 + 0.2 f_3$, where

$$\begin{aligned} f_1 &= MVN(\text{mean}=\mathbf{0}, \text{Cov}=\sum_1'(\rho=0.2, \sigma_{01}^2=1.12, \sigma_{00}^2=0.93, \sigma_{11}^2=1.15, \sigma_{10}^2=0.88)), \\ f_2 &= MVN(\text{mean}=\mathbf{0}, \text{Cov}=\sum_2'(\rho=0.5, \sigma_{01}^2=1.18, \sigma_{00}^2=0.87, \sigma_{11}^2=1.08, \sigma_{10}^2=0.92)) \text{ and} \\ f_3 &= MVN(\text{mean}=\mathbf{0}, \text{Cov}=\sum_3'(\rho=0.7, \sigma_{01}^2=1.24, \sigma_{00}^2=1.06, \sigma_{11}^2=1.27, \sigma_{10}^2=1.05)). \end{aligned}$$

We simulate data with size $n=1,000$ to fit the model and further simulate $n=2,000$ from the fitted model to assess the performance of the proposed approach. For each generated data set (from the fitted model), we fit the normal random effects model and the DPM random effects model. Prior choices and convergence of the Markov chains are assessed exactly as before. MSE is calculated for each simulated data set. We calculate average MSE (AMSE) using 50 such generated datasets. For the normal random effects model we get 2.15 as the AMSE with 1.71 as the standard error. For the DPM random effects model, the AMSE is 0.97 with standard error 0.13. We notice that DPM random effects model performs reasonably better than the normal random effect model, which indicates the superiority of DPM random effects model over the traditional models.

In Table 3, we further demonstrate the superiority of the DPM random effects model. We provide the true values of the model parameters and the width of the estimated 95% credible intervals and the estimated coverage probabilities on the basis of the above simulation. Table 3 shows that DPM random effects model provides much shorter credible intervals with comparable coverage probabilities for the mean parameters. Results for the covariance parameters are given in S.2 in web appendix. Thus the simulation study provides evidence for the practical usefulness of the DPM random effects model.

5 Discussion

We have described a novel approach of jointly modeling the mean and covariance function when the longitudinal response variable is bivariate in nature. The two responses from the same subject are correlated over time which might not be completely explained by random effects. For realistic and meaningful inference, a truncated DPM prior is considered for the random effects distribution. We also propose a nonparametric flexible approach of modeling the bivariate longitudinal covariance matrix based on modified Cholesky decomposition.

More recently, Li et al. [12] proposed semi-parametric mixed models for analyzing longitudinal data in a Bayesian framework. For analyzing univariate longitudinal data, they proposed a mixed model where the mean effect of time was modeled by smoothing splines and a Dirichlet Process (DP) prior was taken for subject-specific random effects. However, our approach differs from theirs in a number of ways. First, we focus on the analysis of bivariate longitudinal traits and jointly model the mean and covariance in a Bayesian framework. Specifically, we don't assume that the residuals in model (1) are independently distributed over time. We propose a flexible approach to model the complex covariance structure for irregular sparse bivariate longitudinal measurements. Second, we consider a DPM of normal prior (not a DP prior) for the random effects distribution for flexibility. Also, we avoid the complexities of smoothing splines for high dimensional data by using penalized splines for estimating $f_k(\cdot)$, the unknown smooth function of time.

In the current work, although we consider subject-specific measurements, we assume two response features are measured simultaneously at each time point for each subject. For traits like systolic and diastolic BP this is true, but this might not be the case in other situations. Especially when extending this approach to multivariate response, it is quite likely that some responses were missing at some time points. However, if the missingness is ignorable (e.g., missing at random), one can simply add a data-augmentation step to the proposed algorithm. For non-ignorable missingness, a more complex method is required (Daniels and Hogan [25]). Addressing the complexities for multivariate sparse longitudinal data will be a possible extension of the proposed method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Li was supported by an NIDA grant P50-DA10075-16 and a NNSF of China grant 11028103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA, or the NIH.

The Framingham Heart Study was conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (N01 HC25195). The Genetic Analysis Workshop is supported by NIH grant R01 GM31575. The simulated data was supported by the Washington University Institute of Clinical and Translational Sciences, NIH grant 1U54RR023496. The GAW16 Framingham and simulated data used for the analyses described in this manuscript were obtained through dbGaP (accession numbers). The authors acknowledge the investigators that contributed data for this study. This manuscript was not prepared in collaboration with

investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI.

References

1. Pan J, Mackenzie G. On modelling mean-covariance structures in longitudinal studies. *Biometrika*. 2003; 90:239–244.
2. Das K, Li J, Fu G, Wang Z, Li R, Wu R. Dynamic semiparametric Bayesian models for genetic mapping of complex trait with irregular longitudinal data. *Statistics in Medicine*. 2013; 32(3):509–523. [PubMed: 22903809]
3. Laird NM, Ware JH. Random effects model for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
4. Zhang D, Lin X, Raj J, Sowers MF. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*. 1998; 93:710–719.
5. Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*. 1994; 8:81–89.
6. Ryu R, Li E, Mallick BK. Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. *Biometrics*. 2010; 10.1111/j.1541-0420.2010.01489.x
7. Müller P, Quintana FA, Rosner GL. Semiparametric Bayesian inference for multilevel repeated measurement data. *Biometrics*. 2007; 63:280–289. [PubMed: 17447954]
8. Ghosh P, Hanson T. A Semiparametric Bayesian approach to multivariate longitudinal data. *Australian and New Zealand Journal of Statistics*. 2010; 52:275–288. [PubMed: 21731424]
9. Bush CA, MacEachern SN. A semiparametric Bayesian model for randomised block designs. *Biometrika*. 1996; 83:275–286.
10. Kleinman KP, Ibrahim JG. A semiparametric Bayesian approach to the random effects model. *Biometrics*. 1998; 54:921–938. [PubMed: 9750242]
11. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica*. 1994; 4:639–650.
12. Li Y, Lin X, Müller P. Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics*. 2010; 66:70–78. [PubMed: 19432777]
13. Ishwaran H, James LF. Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*. 2002; 11:1–26.
14. Das K, Li J, Fu G, Wang Z, Wu R. Genome-wide association studies for bivariate sparse longitudinal data. *Human Heredity*. 2011; 72:110–120. [PubMed: 21996601]
15. Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*. 1999; 86:677–690.
16. Pourahmadi M. Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*. 2000; 87:425–435.
17. Daniels M, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*. 2002; 89:553–566.
18. Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Cambridge University Press; New York: 2003.
19. Blackwell D, MacQueen JB. Ferguson distributions via Polya urn schemes. *Annals of Statistics*. 1973; 1:353–355.
20. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*. 1973; 1:209–230.
21. Ishwaran H, Zarepore M. Markov chain Monte Carlo in approximate Dirichlet and beta two parameter process hierarchical models. *Biometrika*. 2000; 87:371–390.
22. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*. 1998; 7:434–455.
23. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package. *Journal of Statistical Software*. 2010; 33:1–22. [PubMed: 20808728]

24. Sithole JS, Jones PW. Bivariate longitudinal model for detecting prescribing change in two drugs simultaneously with correlated errors. *Journal of Applied Statistics*. 2007; 34:339–352.
25. Daniels MJ, Hogan JW. *Missing data in longitudinal studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall (CRC Press). 2008

Figure 1.
Systolic blood pressures for 10 randomly selected subjects

Figure 2.
Fitted curve for systolic blood pressure when error covariance matrix is modeled as Identity (dotted) and the proposed approach (solid)

Figure 3.
Fitted curve for systolic blood pressure when random effect is modeled by Normal (dotted) and Truncated DPM (solid)

Figure 4.
Fitted curves for the systolic and diastolic pressure

Figure 5.
Posterior density plot for

Table 1

DIC values for selecting the optimum triple (r, g, h) for analysing FHS data.

Triple	DIC
(3,3,3)	175.83
(3,2,3)	163.92
(3,3,2)	181.59
(3,2,2)	158.47
(2,2,2)	138.55
(2,3,2)	146.83
(2,2,3)	144.76
(2,3,3)	151.61

Table 2

RSS and DIC values for different covariance structures for analysing FHS data.

Covariance Structure	Random Effects	RSS	DIC
Independent error	DPM	18.93	229.17
$UN \quad AR(1)$	DPM	16.86	221.89
$UN \quad CS$	DPM	15.74	223.19
$CS + AR(1)$	DPM	15.03	218.75
Proposed approach	DPM	11.98	213.86
Independent error	Normal	20.17	238.72
$UN \quad AR(1)$	Normal	18.93	236.55
$UN \quad CS$	Normal	18.53	225.62
$CS + AR(1)$	Normal	17.85	230.49
Proposed approach	Normal	13.72	218.21

Table 3

Simulation results for the mean parameters to compare the normal random effects model and DPM random effects model in terms of the width of 95% C.I. and the estimated coverage probabilities.

Parameter	Normal r.e. model		DPM r.e. model	
	Width of C.I.	Coverage prob.	Width of C.I.	Coverage prob.
α_1	1.278	0.94	0.173	0.93
α_{11}	0.268	0.94	0.132	0.94
α_{21}	0.271	0.95	0.166	0.94
b_{10}	4.38	0.94	2.04	0.93
b_{11}	0.294	0.95	0.112	0.93
b_{12}	3.228	0.94	1.917	0.93
c_{11}	0.0302	0.96	0.0234	0.94
c_{12}	0.0206	0.95	0.0109	0.95
c_{13}	0.0044	0.94	0.0013	0.93
c_{14}	0.0182	0.95	0.0069	0.94
c_{15}	0.149	0.94	0.108	0.93
α_0	0.231	0.96	0.144	0.95
α_{10}	0.292	0.95	0.176	0.94
α_{20}	0.04	0.94	0.017	0.93
b_{00}	3.84	0.95	2.05	0.94
b_{01}	0.392	0.94	0.179	0.94
b_{02}	0.388	0.94	0.229	0.93
c_{01}	0.0397	0.95	0.0216	0.94
c_{02}	0.0439	0.94	0.0282	0.93
c_{03}	0.0288	0.95	0.0144	0.93
c_{04}	0.0051	0.94	0.0029	0.92
c_{05}	0.0068	0.94	0.0027	0.93