Research Article

Statistics in Medicine

Received 18 February 2010, Accepted 24 January 2011 Published online 13 April 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4236

Flexible estimation of covariance function by penalized spline with application to longitudinal family data

Yuanjia Wang[*†]

Longitudinal data are routinely collected in biomedical research studies. A natural model describing longitudinal data decomposes an individual's outcome as the sum of a population mean function and random subject-specific deviations. When parametric assumptions are too restrictive, methods modeling the population mean function and the random subject-specific functions nonparametrically are in demand. In some applications, it is desirable to estimate a covariance function of random subject-specific deviations. In this work, flexible yet computationally efficient methods are developed for a general class of semiparametric mixed effects models, where the functional forms of the population mean and the subject-specific curves are unspecified. We estimate nonparametric components of the model by penalized spline (P-spline, *Biometrics* 2001; 57:253–259), and reparameterize the random curve covariance function by a modified Cholesky decomposition (*Biometrics* 2002; 58:121–128) which allows for unconstrained estimation of a positive-semidefinite matrix. To provide smooth estimates, we penalize roughness of fitted curves and derive closed-form solutions in the maximization step of an EM algorithm. In addition, we present models and methods for longitudinal family data where subjects in a family are correlated and we decompose the covariance function into a subject-level source and observation-level source. We apply these methods to the multi-level Framingham Heart Study data to estimate age-specific heritability of systolic blood pressure nonparametrically. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: multi-level functional data; Cholesky decomposition; age-specific heritability; Framingham Heart Study

1. Introduction

When longitudinal data are available, nonparametric statistical methods on modeling population mean function and subject-specific functions have been proposed. In Rice and Wu [1], regression splines were used to model random subject-specific curves in a mixed effects model framework. The performance of regression spline is sensitive to the number and location of knots and having a good criterion to choose the number of knots is critical. In Guo [2], functional mixed effects model was considered and a computational intensive Kalman filtering algorithm was introduced to fit model by smoothing splines. To alleviate computational burden, Durban *et al.* [3] pursued a simple and flexible approach to fit subject-specific curves by penalized spline. Specifically, they expressed the subject-specific curves as linear combinations of truncated polynomial spline basis with random coefficients and specified an independent covariance matrix for the coefficients of knots. Owing to the simple form of the covariance matrix, this method allows for fast computation of subject-specific curves. Nevertheless, the independent constraint on the basis coefficient covariance causes the estimated covariance function to be non-invariant to change of the chosen spline basis [4].

In some applications, modeling a covariance function of subject-specific curves is of scientific interest. In other applications, although covariance function itself may not be of direct scientific interest, its

Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, U.S.A.
[*]*Correspondence to: Yuanjia Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, U.S.A.*
[†]*E-mail: yw2016@columbia.edu*

accurate estimation leads to an efficiency gain in estimating population mean function and fixed effects parameters. An example of the variance–covariance function being scientifically interesting is in genetic studies where the covariance function of related subjects in a family represents genetic information as in the motivating example of this work, the Framingham Heart Study (FHS, [5]). The FHS is a large ongoing prospective longitudinal study of risk factors for cardiovascular disease (CVD) first originated in 1948. Discovering genetic risk factors for CVD is one of the major goals in the FHS. The FHS recruits several generations of subjects in families that provide a rich resource for genetic studies. The study collects longitudinal phenotypes such as cholesterol level, blood pressure, and blood glucose.

The existing analysis of genetic studies with longitudinal phenotypes includes stratified analysis which performs a separate genetic analysis for each age stratum [6, 7], two-step-based methods which run genetic analysis on the summary statistics of longitudinal phenotypes [8], and parametric methods which assume a parametric form of the unknown genetic effect [9]. A survey of related methods can be found in Gauderman and Conti [10], Almasy *et al.* [11] and MacCluer *et al.* [12]. Stratified analysis and two-step analysis do not optimally use information in the repeated phenotypes leading to loss of power and parametric analysis being subject to misspecification of the unknown genetic effect. For more flexible methods, Fang and Wang [13] provided the estimation of age-specific heritability using regression splines.

The challenges inherent in analyzing longitudinal family data include dealing with multi-level of correlations: subjects in the same families are correlated and repeated measures on the same subject contribute a second level of correlation. Both levels of correlation need to be accounted for to achieve accurate statistical estimation. In this work, we model the covariance function of genetic effect over time as a Kronecker product of two sources: the between-subject genetic correlation source and the within-subject serial correlation source. The subject-level covariance is predicted by relationship between family members or observed genotypes at genetic markers, whereas the time-level covariance involves an unknown covariance matrix to be estimated. Through penalized splines [14], we propose semiparametric methods to fit both independent subject data and multi-level family data to obtain flexible covariance function and age-dependent heritability without restrictive parametric assumptions.

In this section, we introduce functional mixed effects models for independent data and in the subsequent section we introduce flexible models for family-based data. We present a class of models where population-level curve, subject-level curves, and covariance function of the random subject-specific curves have unspecified functional form. Let i be the index subject and j be the index measurements within a subject. Consider a functional mixed effects model

$$y_{ij} = \mu(t_{ij}) + x'_{ij}\beta + \eta_i(t_{ij}) + c'_i b_i + \varepsilon_{ij}, \tag{1}$$

where $\mu(t)$ is a population mean function, x_{ij} is a vector of covariates for the fixed effects, β is its coefficients, $\eta_i(t)$ is a nonparametric random subject-specific curve, b_i is a vector of parametric subject-specific random effects, c_i is its design vector, and ε_{ij} is a measurement error. Assume that $\eta_i(t)$, b_i, and ε_{ij} are independent and follow

$$\eta_i(t) \sim \mathrm{GP}(0, \varphi), \quad b_i \sim \mathrm{N}(0, \sigma^2 D), \quad \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2),$$

where $\mathrm{GP}(0, \varphi)$ is a Gaussian process with covariance function $\varphi(s, t)$. For simplicity, here we assume β does not depend on time. It is easy to extend the model to include varying-coefficient by penalized spline (see, for example, Chen and Wang [15]). This model consists of four main components: the parametric fixed effects $x'_{ij}\beta$, the nonparametric population mean $\mu(t)$, the parametric random effects $c'_i b_i$, and the functional random effects $\eta_i(t)$. We focus on estimating $\psi(s, t)$ and predicting $\eta_i(t)$ nonparametrically.

We apply penalized spline [14] to estimate nonparametric components of the model. We consider general unstructured covariance matrix for spline basis coefficients and apply a modified Cholesky decomposition [2] to turn a constrained maximization problem to an unconstrained one, and obtain explicit solutions in the maximization step of an EM algorithm. To provide smooth estimates, we penalize the log likelihood by roughness of the fitted functions. The proposed approaches address several limitations of existing methods including parameterizing covariance of random basis coefficients of subject-specific curves [3], requiring small number of knots with regression splines [1], and intensive computation (Kalman filtering) [2]. We also present methods to account for multi-level family data where subjects are correlated, and apply these methods to analyze the FHS systolic blood pressure (SBP) data which reveal the temporal pattern of heritability of SBP.

2. Methods

2.1. Models for family-based genetic studies with longitudinal outcomes

Model (1) can be extended to accommodate multi-level family data. For a family study, such as FHS, the data structure has three levels with subjects nested in a family and observations nested in a subject. Let i be the index family, j be the index subjects within a family, k be the index observations within a subject, t_{ijk} represent the age of a subject at a visit, and n_i denote the number of subjects in a family. A functional mixed effects model for family data is

$$y_{ijk}(t_{ijk})=\mu(t_{ijk})+\alpha_i+x_{ijk}^{\mathrm{T}}\beta+\eta_{ij}(t_{ijk})+\varepsilon_{ijk}(t_{ijk}), \tag{2}$$

where $\mu(t)$ is a population mean function, $\alpha_i \sim \mathrm{N}(0,\sigma_\alpha^2)$ is a random family-specific shared environmental effect, $\eta_{ij}(t)$ is a random subject-specific genetic effect (polygenic effect), x_{ijk} is a vector of covariates such as gender, and $\varepsilon_{ijk}(t)\sim \mathrm{N}(0,\sigma_\varepsilon^2)$ is a residual measurement error. We assume that α_i, η_{ij}, and ε_{ijk} are independent.

In model (2), the function $\mu(t)$ represents how population mean SBP changes with age, x_{ijk} can be time-varying covariates such as whether a subject is receiving an anti-hypertensive treatment at a visit which may change over the course of study, shared environmental effect α_i can be diet shared among family members, and $\eta_{ij}(t)$ represents the random polygenic effect of interest. The variance of the polygenic effect and the total variance of the outcome can be age-dependent, and their ratio is the age-specific heritability. In Section 4, we provide details on age-dependent heritability and illustrate how to estimate it nonparametrically.

2.2. Basis expansion of nonparametric functions

For simplicity, we illustrate our methods through truncated polynomial basis. Extension to other basis such as B-splines is discussed in Section 5. Let $\Theta_{pq}(t)$ denote the vector of pth-order truncated polynomial basis with q knots, that is, $\Theta_{pq}(t)=(1,t,\ldots,t^p,(t-\tau_1)_+^p,\ldots,(t-\tau_q)_+^p)'$, where $\tau_1,\ldots,\tau_q$ is a sequence of knots. Assume that the mean function can be approximated by a linear combination of spline basis, that is

$$\mu(t)=\Theta'_{p_1q_1}(t)\mu,$$

where $\mu=(\mu_0,\ldots,\mu_{p_1+q_1})'$ are unknown basis coefficients. Similarly, assume that the random subject-specific curves can be approximated as

$$\eta_i(t)=\Theta'_{p_2q_2}(t)\eta_i, \tag{3}$$

where $\eta_i=(\eta_{i,0},\ldots,\eta_{i,p_2+q_2})'$ are random basis coefficients distributed as

$$\eta_{i,0},\ldots,\eta_{i,p_2+q_2}\sim \mathrm{N}(0,\sigma^2\Psi).$$

Note that the basis for the population mean function and the subject-specific functions can be different. By (3), the variance–covariance function of the subject-specific curves is then

$$\varphi(s,t)=\sigma^2\Theta'_{p_2q_2}(s)\Psi\Theta_{p_2q_2}(t). \tag{4}$$

The dimension of the covariance matrix Ψ increases with the dimension of the basis $\Theta_{p_2q_2}$.

With the expansions (3), we may write the model (1) in matrix form as

$$\begin{gathered} Y_i=X_i\beta+B_i\mu+C_ib_i+Z_i\eta_i+\varepsilon_i,\\ b_i\sim\mathrm{N}(0,\sigma^2D),\quad \eta_i\sim\mathrm{N}(0,\sigma^2\Psi),\quad \varepsilon_i\sim\mathrm{N}(0,\sigma^2I_{n_i}),\end{gathered} \tag{5}$$

where Y_i is a vector of the ith subject's outcomes, X_i is this person's design matrix of the parametric fixed effects, μ is a vector of basis coefficients for the mean function, $B_i=(\Theta_{p_1q_1}(t_{i1}),\ldots,\Theta_{p_1q_1}(t_{in_i}))'$ is its design matrix consists of the basis function for the population mean function, b_i is a vector of parametric random effects, C_i is its design matrix, η_i is a vector of random basis coefficients for the nonparametric random effects, and $Z_i=(\Theta_{p_2q_2}(t_{i1}),\ldots,\Theta_{p_2q_2}(t_{in_i}))'$ is its design matrix. Here, the estimation of the covariance matrix D for the parametric random effects is a low-dimensional problem, but the estimation of the covariance matrix Ψ for the random effects basis coefficients can be high dimensional.

2.3. Modified Cholesky decomposition

To provide positive-definite estimate of Ψ, we transform a constraint maximization problem to an unconstrained problem by a modified Cholesky decomposition [16]. In Chen and Dunson [16], a positive-definite matrix Ψ of a random effects covariance was decomposed as

$$\Psi = \Lambda\Gamma\Gamma'\Lambda, \tag{6}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$ is a diagonal matrix and $\Gamma = (\gamma_{st})$ is a lower triangular matrix with diagonal elements one. It can be shown that λ_s^2 are the prediction variances of random basis coefficients and γ_{st} are related to the correlation between the basis coefficients. Specifically [17]

$$\lambda_s^2 = \mathrm{var}(\eta_{is}) \quad \text{and} \quad \mathrm{corr}(\eta_{is}, \eta_{it}) = \frac{\sum_{k=1}^{s\wedge t} \gamma_{sk}\gamma_{tk}}{\sqrt{\sum_{k=1}^{s} \gamma_{sk}^2 \sum_{k=1}^{t} \gamma_{tk}^2}}. \tag{7}$$

With decomposition (6), we further reparameterize our model by letting

$$\eta_i = \Lambda\Gamma\xi_i \tag{8}$$

so that the model (5) becomes

$$\begin{aligned} &Y_i = X_i\beta + B_i\mu + C_i b_i + Z_i\Lambda\Gamma\xi_i + \varepsilon_i \\ &\text{and } b_i \sim \mathrm{N}(0, \sigma^2 D), \quad \xi_i \sim \mathrm{N}(0, \sigma^2 I_M), \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2 I_{n_i}). \end{aligned} \tag{9}$$

Note that in (9), the covariance parameters Λ and Γ in Ψ have been turned into 'mean' parameters, allowing for easy computation of the closed-form solutions in the maximization step of the EM algorithm introduced in the following section. Let $\lambda = (\lambda_1, \ldots, \lambda_M)'$ and $\gamma = (\gamma_{ls},\ l = 1, \ldots, M-1;\ s = 1, \ldots, l-1)$ denote the $M(M-1)/2$ free parameters in Γ. With the modified Cholesky decomposition (6), the covariance matrix of the random basis coefficients Ψ is now expressed by free parameters λ and γ without the positive-semidefinite constraint.

2.4. Estimation through penalized splines

For illustrative purpose, we focus on discussing the methods where the parametric random effects are absent, that is, the model

$$y_i(t_{ij}) = \mu(t_{ij}) + X_{ij}\beta + \eta_i(t_{ij}) + \varepsilon_{ij}. \tag{10}$$

Extension to including the parametric random effects is deferred to Section 5. There is a large body of literature on fitting mixed effects models by EM algorithm [18]. In this work, to estimate the large numbers of parameters involved in Ψ while avoiding overfitting and providing smooth fit, we penalized the likelihood by the roughness of the fitted curves.

First we discuss estimation of the population mean function. Let $\psi = (\beta, \mu, \lambda, \gamma, \sigma^2)$ denote all parameters. Dropping constant terms and for given variance component parameters, the penalized marginal log likelihood of Y under the model (10) is

$$l(Y;\psi) = -\frac{1}{2}\sum_i [\log|V_i| + (Y_i - B_i\mu - X_i\beta)' V_i^{-1}(Y_i - B_i\mu - X_i\beta)] - \frac{v_1}{2}\mu' L_1 \mu, \tag{11}$$

where $V_i = \sigma^2(Z_i\Lambda\Gamma\Gamma'\Lambda Z_i' + I_{n_i})$, v_1 is the smoothing parameter for the mean function, and L_1 is a known penalty matrix related to the basis chosen for the mean function. For truncated polynomial basis Θ_{p1q1}, L_1 is a diagonal matrix, $\mathrm{diag}(\mathbf{0}_{p_1+1}, \mathbf{1}_{q1})$, which implies that only the spline coefficients

of the knots $\{\mu_{p_1+1}, \ldots, \mu_{p_1+q_1}\}$ are penalized. The solution to this problem takes the form of a ridge regression estimator

$$(\hat{\beta}', \hat{\mu}')' = \left(\sum_i \tilde{B}_i' V_i^{-1} \tilde{B}_i + v_1 \tilde{L}_1\right)^{-1} \sum_i \tilde{B}_i V_i^{-1} Y_i, \tag{12}$$

where $\tilde{B}_i = (X_i, B_i)$, $\tilde{L}_1 = \text{diag}(\mathbf{0}_{m+p_1+1}, \mathbf{1}_{q_1})$, and m is the dimension of the fixed effects parameter β.

To fit the covariance function, we can treat the random effects ξ_i as unobserved missing data and employ EM algorithm. Ignoring constant terms, the joint complete data likelihood of Y and ξ is

$$l(Y, \xi; \psi) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i [||Y_i - B_i \mu - X_i \beta - Z_i \Lambda \Gamma \xi_i||^2 + \xi_i' \xi_i]. \tag{13}$$

In [19], L_1 or L_2 penalty was added to the log likelihood to estimate a large covariance matrix for balanced data. Here, we penalize roughness of the random subject-specific curves (3) to provide smooth fitting. This penalization is motivated from the idea that the deviations of the subject-specific trajectories from the population mean are realizations of a Gaussian process with a smooth covariance function. The penalized likelihood for estimating variance components is

$$l^p(Y, \xi; \psi) = l(Y, \xi; \psi) - \frac{v_2}{2} \lambda' L_2 \lambda - \frac{v_3}{2} \gamma' L_3 \gamma, \tag{14}$$

where v_2 and v_3 are smoothing parameters, and L_2 and L_3 are penalty matrices. By straightforward matrix algebra, from (8) we obtain the basis coefficients of the fitted subject-specific curves as $\eta_{il} = \sum_{k<l} \lambda_l \gamma_{kl} \xi_k + \lambda_l \xi_l$, $l = p_2+2, \ldots, p_2+q_2+1$, where p_2 is the order of the spline basis used to fit $\eta_i(t)$ and q_2 is the number of knots. For truncated polynomial basis, the roughness of the random subject-specific curve is measured by the sum of the squared knots coefficients, $\sum_{l=1}^{q_2} \eta_{i,p_2+l}^2$. Therefore, the first p_2+1 elements in λ are not involved in measuring the roughness of the fitted curves, which suggests L_2 to be a diagonal matrix $\text{diag}(\mathbf{0}_{p_2+1}, \mathbf{1}_{q_2})$. Similarly, the first $p_2(p_2-1)/2$ elements in γ (when γ is arranged by columns of Γ) are not involved in the roughness of $\Theta_{p_2 q_2}'(t)\hat{\eta}_i$, so that the matrix L_3 is a diagonal matrix with the first $p_2(p_2-1)/2$ diagonal elements zero and the remaining diagonal elements one.

Specifying different smoothing parameters for λ and γ allows for flexibility in modeling the covariance matrix. Recall that λ are proportional to the prediction variances of the random basis coefficients and γ are related to the correlations as shown in (7). It is not guaranteed that the variance and correlation elements of a covariance matrix will have the same smoothness. For example, when v_3 approaches infinity, the covariance matrix for the spline basis coefficients approaches a diagonal matrix so the spline coefficients are independent. When both parameters go to infinity, the random subject-specific curves are approximated by a random polynomial basis. We penalize λ with v_2 to control the effective degrees of freedom of variance parameters and penalize γ with v_3 to control the effective degrees of freedom of correlation parameters. Both λ and γ contribute to fitting subject-specific curves. Huang *et al.* [20] used a similar strategy to smooth the elements in the diagonal matrix and the lower triangular matrix of a Cholesky decomposition separately.

2.5. The EM algorithm

To apply EM algorithm [18] to fit covariance parameters, we first take the conditional expectation of the log likelihood (14) treating random effects as unobserved (E-step) and then maximize this conditional log likelihood to obtain updated parameters (M-step). Note that the conditional mean and variance of the reparameterized random effects are

$$\begin{aligned} \xi_i^{(u+1)} &= E(\xi_i | Y, \psi^{(u)}) \\ &= (\Gamma'^{(u)} \Lambda^{(u)} Z_i' Z_i \Lambda^{(u)} \Gamma^{(u)})^{-1} \Gamma'^{(u)} \Lambda^{(u)} Z_i' (Y_i - B_i \mu^{(u)} - X_i \beta^{(u)}), \end{aligned} \tag{15}$$

$$G_i^{(u+1)} = \text{Var}(\xi_i | Y, \psi^{(u)}) = \sigma^{2(u)} (\Gamma'^{(u)} \Lambda^{(u)} Z_i' Z_i \Lambda^{(u)} \Gamma^{(u)} + I_{n_i})^{-1}, \tag{16}$$

where u is the index iteration. The conditional expectation of the second term in (13) is then given by (E-step)

$$E_{\xi|Y,\psi^{(u)}}\left\{\sigma^{-2}\sum_i ||Y_i - B_i\mu - X_i\beta - Z_i\Lambda\Gamma\xi_i||^2 + \xi_i'\xi_i\right\}, \tag{17}$$

which can be computed based on (15) and (16).

For the M-step, we show in the appendix that by some matrix manipulations, the penalized likelihood (14) can be written as quadratic forms of parameters λ and γ. Therefore, explicit solutions of the maximization step can be obtained. We show the details of the algorithm in the appendix.

2.6. *Confidence bands*

The point-wise confidence band for the population mean function can be computed based on (12) (see for example, 7.61 in [21]). The variability band of the covariance function is less straightforward. In traditional mixed effects models, standard errors of the estimated variance components $\hat{\lambda}$ and $\hat{\gamma}$ can be obtained from the observed information matrix (see, for example [22]). However, the fitted covariance is a function of $\hat{\lambda}$ and $\hat{\gamma}$ through (4) and (6). To compute standard errors based on observed information matrix, delta method would be used. Here instead, we use straightforward parametric bootstrap to obtain confidence band for the estimated covariance function directly. We resample subject-specific random effects from a Gaussian process with fitted covariance function and resample residuals based on (A6).

2.7. *Choosing tuning parameters*

The tuning parameters for penalized spline include the number and location of knots and smoothing parameters v_1, v_2, and v_3. In Ruppert [23], it was shown that when the number of knots is adequately large, further increasing it does not improve fit, and the smoothing parameter plays a critical role. Yu and Ruppert [24] reported using 5–10 knots to be sufficient for smooth functions. In this work, we use a sufficient number of knots (10) and place the knots at equal sample quantiles of t_{ij} as suggested in [24]. There is a large body of literature on how to choose smoothing parameter in smoothing splines and penalized spline. Popular methods include cross-validation, generalized cross-validation, information criterion-based approaches, such as AIC and BIC, and estimating by maximizing a restricted likelihood [25]. To avoid complications, here we select v_1 by treating it as an extra variance component and estimating through restricted maximum likelihood which was shown in Krivobokova and Kauermann [26] to outperform AIC-based choice with correlated data. We select v_2 and v_3 by cross-validation because the link between these smoothing parameters of the random subject-specific curves and variance components in a linear mixed model is not clear.

3. Simulated examples

In this section, we perform two sets of simulation studies to evaluate the performance of the proposed methods. We simulate 200 subjects in each setting and each subject has four observations. The time interval between observations range from 1 to 4 years. We repeat each experiment 200 times. We use 10 knots for the mean function and 10 knots for the functional random effect.

We simulate data from model (10). The population mean function is a sine function, $\mu(t) = 30\sin(2\pi t/30)$. We simulate a binary covariate with an effect of five, and the residual random error has a variance of one. We consider the following two functional random effects:

(1) Example 1: The random effects are simulated from

$$\eta_i(t) = b_{i1} + b_{i2}\eta(t), \quad b_{i1} \sim \mathrm{N}(0, \sigma_1^2), \quad b_{i2} \sim \mathrm{N}(0, \sigma_2^2),$$

where $\eta(t) = \{0.15\exp(0.05t)\}^{1/2}$, $\sigma_1^2 = 4$, $\sigma_2^2 = 2.25$, and b_{i1} and b_{i2} are independent. The residual random error is simulated from a standard normal distribution. In this example, the variance function of the random effect $\eta_i(t)$ is $4 + 0.34\exp(0.05t)$, which increases over time.

(2) Example 2: In this example, we keep everything the same as in example 1 except that $\eta(t) = \{\exp(-0.08t + 3)\}^{1/2}$. The variance function of $\eta_i(t)$ is $4 + 2.25\exp(-0.08t + 3)$, which decreases over time.

Table I. Bias and MASE of estimated functions.

	Example 1		Example 2	
Function	max bias*	MASE($\hat{f}$)	max bias	MASE($\hat{f}$)
$\hat{\mu}(t)$ Unweighted†	0.213	0.224	0.307	0.164
$\hat{\mu}(t)$ Proposed‡	0.141	0.102	0.257	0.094
$\hat{\mu}(t)$ Alternative§	0.114	0.130	0.101	0.130
$\hat{\sigma}^2_\gamma(t)$ Proposed¶	0.275	0.937	0.417	0.901
$\hat{\sigma}^2_\gamma(t)$ Alternative‖	1.414	0.942	2.334	1.393
Estimator	Mean	SE	Mean	SE
$\hat{\beta}$	4.95	0.353	4.95	0.406
$\hat{\sigma}^2$	0.996	0.021	1.001	0.021

*max bias is defined as $\max_t |\text{mean}(\hat{f}(t)) - f(t)|$.
†$\hat{\mu}(t)$ estimated by (10) with V_i assumed to be an independent matrix.
‡$\hat{\mu}(t)$ estimated by (10) with V_i estimated by the proposed $\hat{\gamma}(s,t)$ with Cholesky decomposition.
§$\hat{\mu}(t)$ estimated by (10) with V_i estimated directly without Cholesky decomposition.
¶$\hat{\sigma}^2_\gamma(t)$ with Ψ estimated by the proposed modified Cholesky decomposition.
‖$\hat{\sigma}^2_\gamma(t)$ with Ψ estimated directly without decomposition (6).

We compute the mean average-squared error (MASE) of the estimated mean function defined as the mean across the 200 simulations of the average-squared error

$$\text{ASE}(\hat{\mu}) = \frac{1}{\sum_i n_i} \sum_{ij} \{\hat{\mu}(t_{ij}) - \mu(t_{ij})\}^2.$$

Define the MASE for the estimated variance function in a similar fashion. Table I summarizes the maximal bias and the MASE of all the estimated functions. We compare the MASE of $\hat{\mu}(t)$ assuming an independent covariance for the outcomes (i.e. assuming $V_i = \sigma^2 I_{n_i}$ in (12); label as '$\hat{\mu}(t)$ unweighted' in Table I) with (a): $\hat{\mu}(t)$ obtained with the covariance matrix Ψ estimated from the proposed Cholesky decomposition (i.e. V_i in (12) estimated as proposed in Section 2.4; labeled as '$\hat{\mu}(t)$ proposed' in Table I); and (b): $\hat{\mu}(t)$ obtained with covariance function matrix Ψ estimated directly without applying the Cholesky decomposition (labeled as '$\hat{\mu}(t)$ alternative' in Table I). The improvements in MASE of the proposed method over the unweighted estimates are 54 and 43 per cent, respectively for each example, which are substantial. Compared to estimating Ψ directly without applying the modified Cholesky decomposition (5), the MASE of $\hat{\mu}(t)$ reduced by 30 and 38 per cent, respectively. The mean and standard error of the estimated covariate effects and the residual variance with the proposed methods are shown in Table I. To investigate the performance of the estimated variability band for the mean function, we compare it with the empirical variability band in the left panel of Figure 1. In general, the estimated standard error is close to the empirical standard error. The variability is largest at the extreme values of t. The standard error estimator is slightly conservative at around age 45.

We use two ways to evaluate the performance of the estimated covariance functions. First, we evaluate the estimated variance function since in many applications variance function is of scientific interest. For example, in our real-data example in Section 4, the variance function is used to construct heritability. For a typical run, the smoothing parameter v_2 is chosen by cross-validation as zero and v_3 as 10^5 in the first example, and they are chosen as 10^4 and zero for the second example again by cross-validation. These simulations show the different roles of the two smoothing parameters. Smoothing parameter v_2 controls the effective degrees of freedom of variance parameters and v_3 controls the effective degrees of freedom of correlation parameters. The two smoothing parameters may take different values for an application. The MASE of the variance function is small in both examples (Table I). When compared to estimating Ψ without the Cholesky decomposition, the reduction in MASE is 55 per cent for the second example. We compare the bootstrap standard error of the estimated variance function with the empirical standard error in the right panel of Figure 1. We see that the bootstrap standard error is close to the empirical one and the standard error was largest at the extreme values of t.

Second, as suggested in [1], we consider a summary measure in terms of the increase in prediction error. We use the estimated covariance estimator to construct BLUPs of the random effects at the

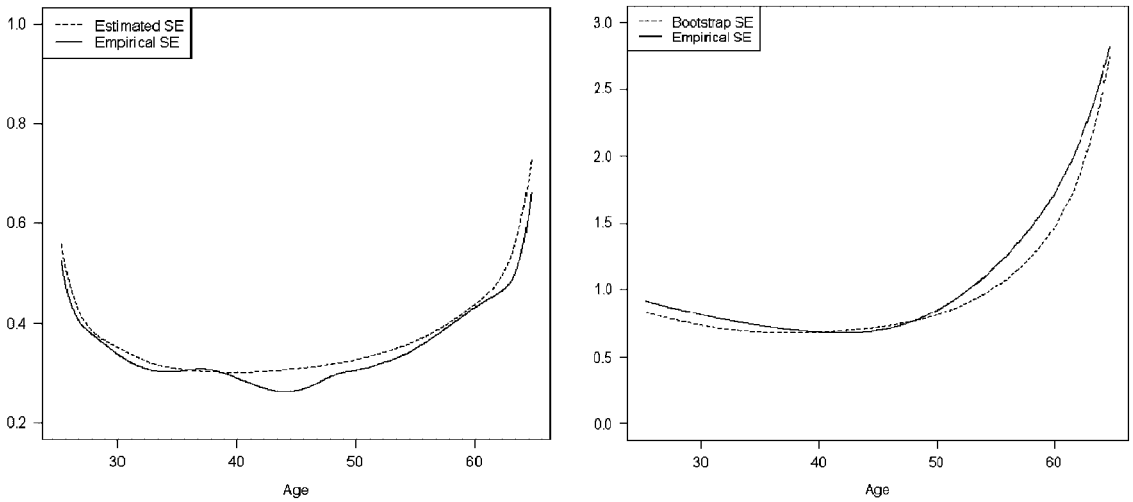

Figure 1. Bootstrap and empirical standard error of the estimated population mean function (left panel) and estimated variance function (right panel).

observed time points and computed the average-squared prediction error across 200 simulation runs as

$$\frac{1}{L}\sum_{l}\sum_{i,j}\{y_{ij}-\hat{\mu}^{(l)}(t_{ij})-X_{ij}\hat{\beta}^{(l)}-\hat{\eta}_i^{(l)}(t_{ij})\}^2,$$

where l is the index each repetition of the simulation and $\hat{\eta}_i$ is the BLUP using estimated covariance matrix. We then compare this prediction error with the best estimate, which uses the true covariance function in predicting random effects BLUPs. The increase in the root mean squared prediction error using the estimated covariance function are 3 and 1.3 per cent for the two settings, respectively, which is minimal.

4. A real-data example

In this section, we apply proposed methods to the longitudinal data collected in the Framingham Heart Study (FHS, [23]) to estimate age-specific heritability of a phenotype nonparametrically. The objective of the FHS is to identify common risk factors or characteristics that contribute to CVD. The study follows development of CVD over a long period of time in a large cohort without overt symptoms of CVD or disease at the time of enrollment. Longitudinal measurements are collected on subjects' clinical characteristics such as cholesterol level, blood pressure, and blood glucose. The FHS recruited several generations of subjects in families: The Original Framingham Cohort (Cohort 1) was first examined in 1948 and has been examined every two years thereafter; The Offspring Cohort (Cohort 2), composed primarily of offspring of the original cohort and the spouses of these offspring, was examined first in 1971 and has been examined approximately every four years.

Among phenotypes collected at the FHS, high blood pressure is a major risk factor for stroke and heart disease and it affects about one-third of the adult population in the U.S. [5]. Systolic and diastolic blood pressure (SBP and DBP) are complex traits that may be influenced by both environmental and genetic factors. The long-term average heritability of systolic blood pressure is estimated to be high (30–60 per cent, [27]), which suggests a substantial genetic contribution.

Although genetic contribution to blood pressure is noted by researchers, age-dependent genetic effect is routinely ignored in genetic analysis, making discovery of individual genes with moderate effects more difficult, potentially leading to inconsistent replication of gene-association findings [28]. Here, we apply proposed methods to estimate age-specific heritability of SBP nonparametrically. Simply speaking, in statistical genetics variability in the outcome is decomposed into a random genetic component and residual environmental component. The heritability is defined as the ratio of the variance of the two components [29]. We attempt to obtain nonparametric estimation of heritability.

A semiparametric model for family-based genetic study with longitudinal outcomes is presented in (2). The nonparametric random polygenic effect $\eta_{ij}(t)$ represents the overall genetic information. This effect is characterized by its covariance function which is related to the relationship between relatives in a family. To be specific, express $\eta_{ij}(t)$ in terms of spline basis as

$$\eta_{ij}(t)=\Theta'(t)\eta_{ij},$$

where $\Theta(t)$ is a q-dimensional vector of truncated polynomial basis and η_{ij} is the corresponding vector of subject-specific polygenic coefficients. The covariance matrix of η_{ij} can be specified as Meyer [30]

$$\mathrm{Cov}(\eta_{ij}^{(l)},\eta_{ij'}^{(l')})=K_{jj'}^{i}\sigma_{ll'}^{2} \quad \text{for } j,j'=1,\ldots,n_i, \quad l,l'=1,\ldots,p_2+q_2+1, \tag{18}$$

where $\eta_{ij}^{(l)}$ is the lth component of the vector η_{ij}, and $\mathrm{Cov}(\eta_{ij}^{(l)},\eta_{i'j'}^{(l')})=0$ for $i\neq i'$. Here, $K_{jj'}^{i}$ denotes the known kinship coefficient between two subjects in a family, and $\Omega=\{(\sigma_{ll'}^{2}),\ l,l'=1,\ldots,p_2+q_2+1\}$ is the unknown covariance matrix of the polygenic effects basis. The kinship coefficient is defined as the probability of randomly drawing an allele in subject j that is identical by descent (IBD) to an allele at the same locus randomly drawn from subject j' [29]. For example, twice the kinship coefficient for a full sibling pair is 1/2 and for a half-sibling pair is 1/4. These coefficients are calculated based on the relationship between subjects in a family.

The specification (18) essentially models the covariance function of random genetic effect over time as a Kronecker product of two sources: the subject-level source and the observation-level source. The subject-level covariance is predicted by the kinship coefficients based on relationship between relatives in a family, while the observation-level covariance is specified by basis vector $\Theta(t)$ and the unknown covariance matrix Ω. For example, for a family with two siblings and each sibling with two measurements, for instance, the covariance matrix is

$$Z_i'(K_i\otimes\Omega)Z_i, \tag{19}$$

where

$$K_i=\begin{pmatrix}1 & 1/2\\ 1/2 & 1\end{pmatrix}, \quad Z_i=\mathrm{diag}(Z_{i1},Z_{i2})$$

and $Z_{ij}=(\Theta(t_{ij1}),\ldots,\Theta(t_{ijn_{ij}}))'$.

The main goal in this analysis is to estimate population mean function and heritability defined as the ratio of the genetic variance and the total trait variance. By model (2) and expression (18), the age-specific heritability is

$$h^2(t)=\sigma_\eta^2(t)/\sigma_T^2(t), \tag{20}$$

where

$$\sigma_\eta^2(t)=\Theta'(t)\Omega\Theta(t) \tag{21}$$

is the genetic variance, and $\sigma_T^2(t)=\mathrm{Var}(Y(t))$ is the total outcome variance (see also [25, 28]). We use the reparameterization in [31] to express the polygenic effect into a few family-specific and subject-specific random effects to analyze the FHS data and we split large pedigrees into sib-ships for fast computation.

We restrict our attention to observations between age 30 and 60 and subjects who did not take anti-hypertensive treatment. There are 2100 observations from 419 subjects and 192 sib-pairs from 147 families. On average, there are 5.01 observations per subject and 2.18 subjects per sib-pair. Figure 2 shows a scatter plot of the SBP against age for all subjects.

We fit model (2) with sex as a fixed effect, and a sib-ship-specific random genetic effect $\gamma_{ij}(t)$. There are 10 knots for the mean function and 10 knots for the random genetic function. The number of knots was chosen following Ruppert [23] and Yu and Ruppert [24]. The fitted functions do not differ when increasing the number of knots from 10 to 20 (maximal difference in $\hat{\mu}(t)$ across time is less than 0.1). Figure 3 shows the population mean curve along with its confidence interval. It can be seen that the mean SBP stays rather flat at around 126.8 mm Hg (CI: 123.6, 129.9) from age 30 to around 37 and starts to increase from age 37. The mean SBP reaches 138.6 mm Hg (CI: 134.8, 142.4) at age 60. The gender effect is estimated as 4.98 (CI: 3.59, 6.37) with men having higher SBP, on average.

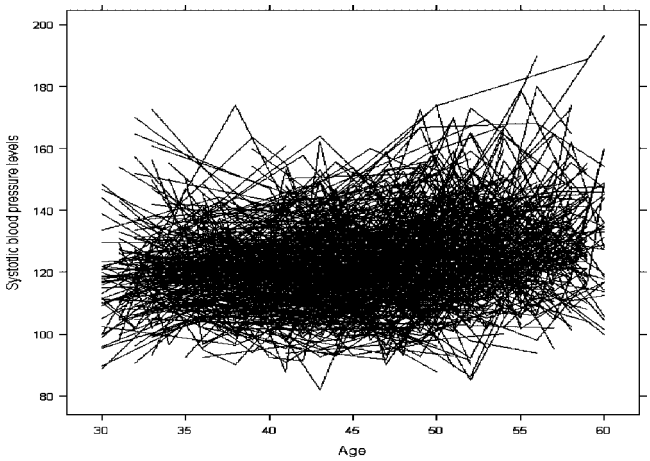

Figure 2. Scatter plot of SBP versus age in the Framingham Heart Study.

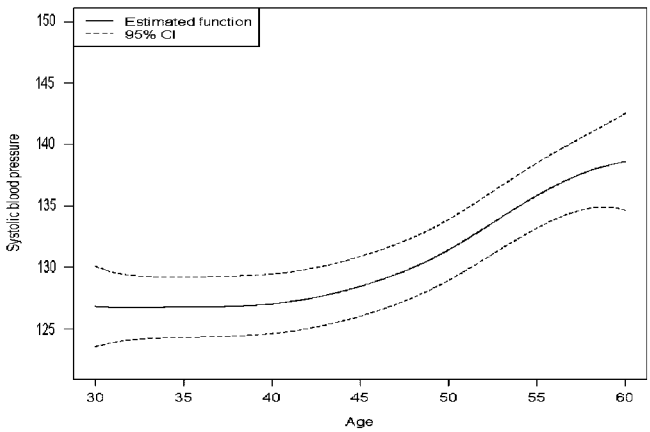

Figure 3. Estimated population mean function of SBP versus age in the Framingham Study.

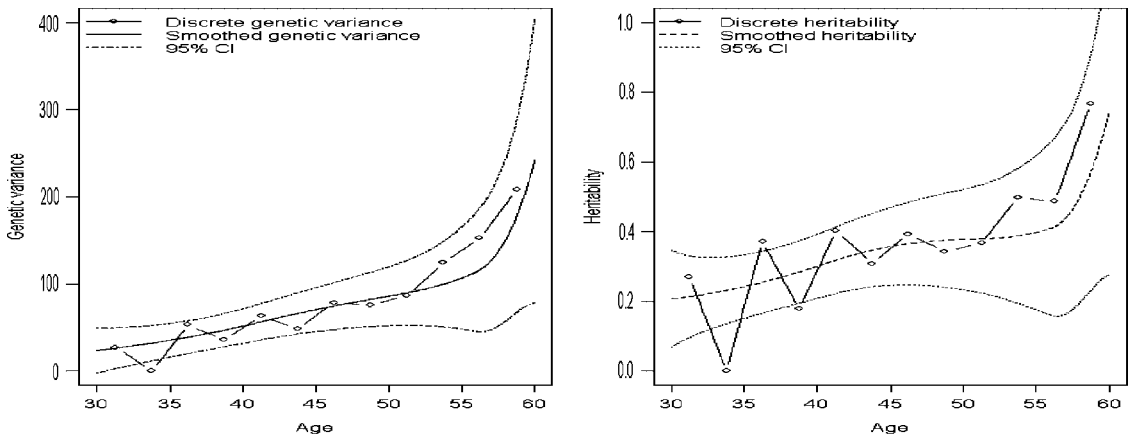

Figure 4. Estimated age-specific genetic variance (left panel) and age-specific heritability (right panel) in the Framingham Heart Study. Smoothed variance and heritability are obtained by penalized splines; discrete variance and heritability are obtained by dividing observations into 12 2.5-year intervals, averaging all repeated measurements in that interval collected on the same subject, and fitting one polygenic model on the averaged observations in each interval.

The left panel in Figure 4 shows the estimated genetic variance function $\sigma_\gamma^2(t)$ as defined in (21) along with its bootstrap confidence interval. The genetic variance increases slowly from age 30 to 57, and the rate of increase is much higher from age 57 to 60. We compare it with a stratified estimate

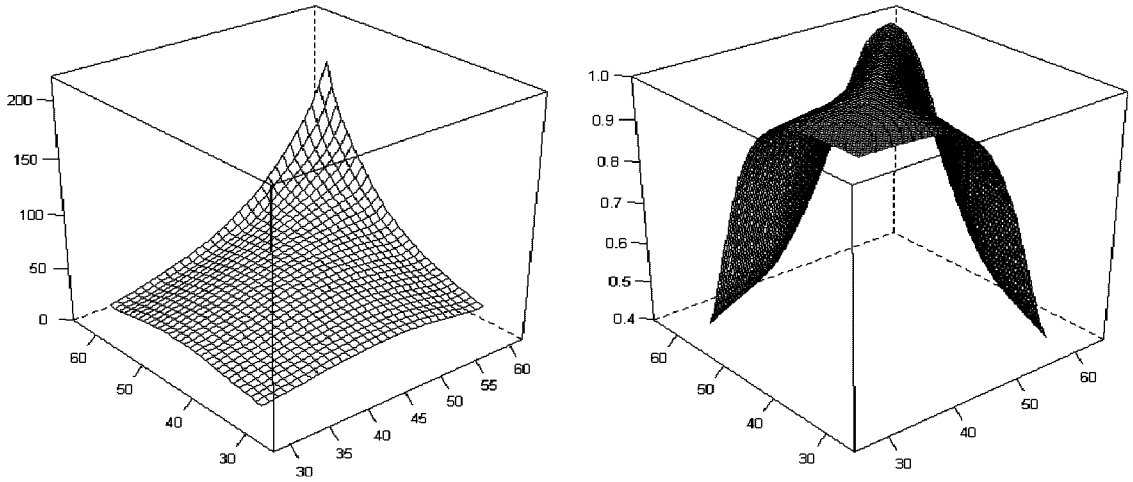

Figure 5. Estimated genetic covariance (left panel) and correlation (right panel) of systolic blood pressure in the Framingham Heart Study.

(discrete estimate) obtained by dividing observations into 12 intervals with 2.5-year width, averaging all repeated measurements in each interval collected on the same subject, and fitting a model on the averaged observations in each interval. It can be seen that the genetic variance estimated by penalized spline serves as a smoothed version of the stratified estimate. The right panel in Figure 4 shows the estimated heritability as defined in (20), where we estimate the total trait variance by a penalized spline smoothing of squared residuals, $[y_{ijk} - x_{ijk}\hat{\beta} - \hat{\mu}(t_{ijk})]^2$. The heritability increases from 0.21 (CI: 0.07, 0.35) at age 30 to 0.37 (CI: 0.18, 0.49) at age 47. It stays rather flat from age 47 to 54 and then enters a period of fast increase before reaching 0.74 (CI: 0.27, 1.00) at age 60. Again we compare it with the stratified estimate obtained by averaging repeated measurements on the same subject in an age stratum. The stratified estimate is rough and highly variable while the penalized spline estimate is smooth. We show the fitted genetic covariance and correlation function for SBP in Figure 5. The correlation between time points decreases as the time separation increases.

As a sensitivity analysis, we fit a parametric quadratic function for the polygenic covariance to examine its effect on estimating heritability. The heritability was estimated as 0.202, 0.311, 0.413, and 0.504 at age 30, 40, 50, and 60, respectively, compared to the nonparametric estimates obtained at the same ages as 0.207, 0.300, 0.377, and 0.744. The parametric estimates captured the general increasing trend, but was not flexible enough to capture the trajectory of the heritability estimates for older ages where there are less observations available.

In Shi *et al.* [32], a parametric model was applied to incorporate age trend to a variance components model in a genetic linkage study with cross-sectional data. The authors assumed a Gaussian function of age for polygenic variance. In their analysis, the polygenic heritability of systolic blood pressure was estimated to peak at age 74.4 with an estimate of 0.69 in Caucasians and peak at age 58.5 with an estimate of 0.68 in African Americans [32]. Here, we take a nonparametric approach without assuming the shape of polygenic variance. The estimated heritability at the older ages in [32] were similar to the estimates obtained in our analysis although the shape of age-specific heritability can differ.

5. Discussion

In this work we propose nonparametric methods to estimate time-varying population mean and covariance function from a collection of random curves by penalized spline. The penalized spline method uses a moderate number of knots without sacrificing quality of fit [23]. The proposed methods are applicable to unbalanced longitudinal data and are easy to implement. We use a Cholesky decomposition of the covariance function to provide unconstrained estimates. Similar Cholesky decomposition for random effect covariance was used in [16, 33] to perform variable selection in linear mixed effects models. The developed methods are applied to estimate age-dependent heritability in the Framingham study where subjects collected in a family are correlated. The heritability estimates are useful for planning a future

study and developing treatment for a disease: sampling subjects at the age where heritability is at its peak would enhance power of an association study; interventions may target different environmental or genetic factors at different ages depending on which factor dominates.

In Section 2, we illustrate our methods assuming the parametric random effects $C_i b_i$ are absent. To extend our methods to handle these effects, in the E-step, in addition to (15) and (16) we update b_i by

$$b_i^{(u+1)} = D^{(u)} C_i' (V_i^{(u)})^{-1} (Y_i - B_i \mu^{(u)} - X_i \beta^{(u)})$$

and in the M-step, in addition to update (A2) and (A4) we update D by

$$D^{(u+1)} = \frac{1}{n} \sum_i \left\{ b_i^{(u+1)} b_i^{'(u+1)} + [D^{(u)} - D^{(u)} Z_i' (V_i^{(u)})^{-1} Z_i D^{(u)}] \right\},$$

where $V_i^{(u)} = \sigma^{2(u)} (C_i D^{(u)} C_i' + Z_i \Lambda^{(u)} \Gamma^{(u)} \Gamma^{'(u)} \Lambda^{(u)} Z_i' + I_{n_i})$.

Although the proposed methods are illustrated through truncated polynomial basis, other basis such as B-splines can also be used. The penalty matrix involved in the log likelihood should be adapted accordingly when using other basis functions. For B-splines, Eilers and Marx [14] proposed a difference-based penalty. In [34], a direct generalization of smoothing splines penalty was considered (O'Sullivan penalized spline) and mixed model representation was provided. These works allow our methods to be easily extended to B-splines.

In some applications, investigators may be interested in testing whether the between-subject covariance function differs between two groups. For instance, in the Framingham example, one may be interested in testing whether the genetic variance differs by gender. A model allowing for group-specific covariance function would then be

$$y_{ij}(t_{ij}) = \mu(t_{ij}) + X_{ij}\beta + \eta_i^{w_i}(t_{ij}) + \varepsilon_{ij},$$

where $w_i = 0,$ or 1 is a group indicator, and $\eta_i^{w_i} \sim \mathrm{GP}(0, \psi^{w_i})$ is a Gaussian process with covariance function ψ^{w_i}. The between group difference in covariance function is tested by $\psi^1 = \psi^0$. A similar extension can be applied to the mean function to accommodate group-specific $\mu(t)$.

Note that in model (1), the residuals are assumed to have a constant variance over time given the random subject-specific curves. A more general model would allow $\varepsilon_i(t)$ to follow a non-stationary Gaussian process [15]. In addition, the proposed methods assume normality of the random effects and the residuals. This assumption can be assessed by Q–Q plots. When the assumption does not hold, robust regression methods such as quantile regression or M-estimators may be considered.

In this work, we split the pedigrees into sib-ships when analyzing the FHS data to speed up computation. The dimension of matrix $K_i \otimes \Omega$ in (19) gets larger when the number of subjects in a family increases and the computation burden becomes heavier especially when using cross-validation to select smoothing parameters. A more efficient algorithm with full pedigree linkage analysis is being developed elsewhere.

Other alternative decompositions of a covariance matrix have been proposed in the literature. For example, authors in Fan *et al.* [35] decomposed the covariance matrix into a variance–correlation form and estimate the variance function nonparametrically but the correlation function parametrically. For this factorization, it is difficult to estimate correlation nonparametrically while satisfying the positive-semidefinite constraint. In [19], a modification of Cholesky decomposition, $T\Sigma T' = D$, was used. The components in T are obtained by regressing Y_t on its predecessors, $Y_1, \ldots, Y_{t-1}$, and the components in D are innovation variances of the regression. However, since the random effects in our model are not observed, no straightforward computation of T and D is available. Spectral decomposition and principal components analysis are other popular methods applied to estimate covariance function [36, 37]. This decomposition does not permit a simple conditional linear form and may require applying a surface smoother to estimate the functional covariance [36] or orthogonalization of an estimated covariance [37]. In addition, the estimated covariance function is not guaranteed to be positive semidefinite [38].

Appendix A

For the M-step in the EM algorithm, for given variance components parameters $\lambda^{(u)}, \gamma^{(u)}$, and $\sigma^{(u)}$, we can solve for coefficients for nonparametric population mean μ by maximizing (17). Setting the

derivative of (17) to be zero to obtain estimating equation for μ as

$$\hat{\mu}^{(u+1)}=\left\{\sum_i[\Theta_i'\Theta_i+v_1L_1]\right\}^{-1}\sum_i\Theta_i'(Y_i-Z_i\Lambda^{(u)}\Gamma^{(u)}\xi_i^{(u+1)}). \tag{A1}$$

We then develop the closed-form solution for the variance components parameters λ. Since Λ is a diagonal matrix, it can be shown that $Z_i\Lambda\Gamma\xi_i=Z_i\mathrm{Diag}(\Gamma\xi_i)\lambda$. Expanding the first term in (17) and dropping terms without λ, we obtain

$$\lambda'\left[\sum_i\mathrm{Diag}(\Gamma\xi_i)Z_i'Z_i\mathrm{Diag}(\Gamma\xi_i)\right]\lambda-2\left[\sum_i(Y_i-\Theta_i'\mu)'Z_i\mathrm{Diag}(\Gamma\xi_i)\right]\lambda+v_2\lambda'L_2\lambda, \tag{A2}$$

which is a quadratic form in λ. By straightforward but tedious matrix algebra, one can show that

$$\mathrm{Diag}(\Gamma\xi_i)Z_i'Z_i\mathrm{Diag}(\Gamma\xi_i)=Z_i'Z_i\circ\Gamma\mathrm{Diag}(\xi_i)\mathbf{1}\mathbf{1}'\mathrm{Diag}(\xi_i)\Gamma=Z_i'Z_i\circ\Gamma\xi_i\xi_i'\Gamma',$$

where '$\circ$' denotes Hadamard (element by element) product (see also [18]). Further observe that

$$E_{\xi_i|Y_i,\psi^{(u)}}(\xi_i\xi_i')=G_i^{(u)}+\xi_i^{(u)}\xi_i^{'(u)}$$

we can then take the conditional expectation of (A2) and set its derivative to zero to obtain the update function for λ. That is

$$\lambda^{(u+1)}$$
$$=\left\{\sum_iZ_i'Z_i\circ\Gamma^{(u)}\xi_i^{(u+1)}\xi_i^{'(u+1)}\Gamma^{'(u)}+v_2L_2\right\}^{-1}\sum_i\{Z_i\mathrm{Diag}(\Gamma^{(u)}\xi_i^{(u+1)})\}'(Y_i-B_i\mu^{(u)}-X_i\beta^{(u)}). \tag{A3}$$

Next, we turn to deal with the covariance parameters γ which are related to the correlation between basis coefficients. When $\gamma=0$, the basis coefficients are independent and Ψ reduces to a diagonal matrix. To write (17) in a quadratic form of γ, observe that the first term of (17) can be re-expressed as [2]

$$\sum_{ij}(Y_{ij}-B_{ij}'\mu-w_{ij}'(\xi_i)\gamma-Z_{ij}'\Lambda\xi_i)^2,$$

where we define the $M(M-1)/2$-dimensional vector $w_{ij}(\xi_i)=(\xi_{il}\lambda_mZ_{ijm}:l=1,\ldots,M,\ m=l+1,\ldots,M)'$. By this expression, we can write the conditional expectation of (17) as a quadratic form in γ, that is

$$\gamma'\sum_{ij}E_{\xi|Y,\psi^{(u)}}(w_{ij}w_{ij}')\gamma-2\sum_{ij}E_{\xi|Y,\psi^{(u)}}[(Y_{ij}-B_{ij}'\mu-Z_{ij}'\Lambda\xi_i)w_{ij}']+v_3\gamma'L_3\gamma.$$

Solve the above equation to obtain the update function for γ as

$$\gamma^{(u+1)}$$
$$=\left\{\sum_{ij}E_{\xi|Y,\psi^{(u)}}w_{ij}^{(u)}w_{ij}^{'(u)}+v_3L_3\right\}^{-1}\sum_{ij}E_{\xi|Y,\psi^{(u)}}\{w_{ij}^{(u)}(Y_{ij}-B_{ij}\mu-X_i\beta^{(u)}-Z_{ij}'\Lambda^{(u)}\xi_i^{(u+1)})\}. \tag{A4}$$

Note that w_{ij} is a function of ξ_i so that the conditional expectation in (A4) involves the first and the second conditional moment of ξ_i which are computed from (15) and (16).

Lastly, the residual variance σ^2 is updated by

$$\sigma^{2(u+1)}=\frac{1}{N}\sum_i\{\hat{\varepsilon}_i^{'(u)}\hat{\varepsilon}_i^{(u)}+\sigma^{2(u)}[n_i-\sigma^{2(u)}\mathrm{Tr}(P_{V_i^{(u)}})]\}, \tag{A5}$$

where

$$\hat{\varepsilon}_i^{(u)}=Y_i-B_i\mu^{(u)}-X_i\beta^{(u)}-Z_i\Lambda^{(u)}\Gamma^{(u)}\xi_i^{(u)} \tag{A6}$$

and $\mathrm{Tr}(P_{V_i^{(u)}})$ is the trace of the projection matrix, $P_{V_i^{(u)}}$.

To summarize, our EM algorithm is

(1) Obtain starting value for ψ^0 by assuming a working independent correlation matrix for random basis coefficients η_i.
(2) For given $\Lambda^{(u)}$, $\Gamma^{(u)}$, and $\sigma^{2(u)}$, update μ and β by (A1).
(3) For given $\mu^{(u+1)}$, $\beta^{(u+1)}$, $\Gamma^{(u)}$, and $\sigma^{2(u)}$, update ξ_i and G_i by (15) and (16) and then update Λ by (A3).
(4) For given $\mu^{(u+1)}$, $\beta^{(u+1)}$, $\Lambda^{(u+1)}$ and $\sigma^{2(u)}$, update ξ_i and G_i by (15) and (16) and then update Γ by (A4).
(5) For given $\mu^{(u+1)}$, $\beta^{(u+1)}$, $\Lambda^{(u+1)}$, and $\Gamma^{(u+1)}$, update σ^2 by (A5).
(6) Iterate steps 2 through 5 until the final convergence is reached.

We implement the algorithm as in R programming language (http://cran.r-project.org/). The core codes can be downloaded from www.columbia.edu/∼yw2016/CovSel.R.

Acknowledgements

The data were obtained from the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine (Contract No. N01-HC-25195). Yuanjia Wang's research is supported by NIH grant AG031113-01A2. The author thanks Dr Chiahui Huang for assisting with computer programming.

References

1. Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 2001; **57**:253–259.
2. Guo W. Functional mixed effects models. *Biometrics* 2002; **58**:121–128.
3. Durban M, Harezlak J, Wand MP, Carroll RJ. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine* 2005; **24**:1153–1167.
4. White IMS, Cullis BR, Gilmourr AR, Thompson R. Smoothing biological data with splines. *Proceedings of International Biometrics Conference*, Cape Town, South Africa, December 1998.
5. Dawber TR, Meadors GF, Moore FEJ. Epidemiological approaches to heart disease: the Framingham study. *American Journal of Public Health* 1951; **41**:279–286.
6. Rodríguez-Rodríguez E, Infante J, Llorca J, Mateo I, Sánchez-Quintana C, García-Gorostiaga I, Sánchez-Juan P, Berciano J, Combarros O. Age-dependent association of KIBRA genetic variation and Alzheimer's disease risk. *Neurobiol Aging* 2009; **30**(2):322–324.
7. Zerba KE, Ferrell RE, Sing CF. Genotype–environment interaction: apolipoprotein E (ApoE) gene effects and age as an index of time and spatial context in the human. *Genetics* 1996; **143**:463–478.
8. Strauch J, Golla A, Wilcox MA, Baur MP. Genetic analysis of phenotypes derived from longitudinal data: presentation group 1 of Genetic Analysis Workshop 13. *Genetic Epidemiology* 2003; **25**(Suppl. 1):S5–S17.
9. Shi G, Rao DC. Ignoring temporal trends in genetic effects substantially reduces power of quantitative trait linkage analysis. *Genetic Epidemiology* 2008; **32**:61–72.
10. Gauderman WJ, Conti D. Commentary: models for longitudinal family data. *International Journal of Epidemiology* 2005; **34**:1077–1079.
11. Almasy L, Amos CI, Bailey-Wilson JE, Cantor RM, Jaquish CE, Martinez M, Neuman RJ, Olson JM, Palmer LJ, Rich SS, Spence MA, MacCluer JW. Genetic analysis workshop 13: analysis of longitudinal family data for complex diseases and related risk factors. *BMC Genetics* **4**(Suppl. 1):S1.
12. MacCluer JW, Cupples LA, Almasy L. Genetic analysis workshop 16: approaches to analysis of genome-wide data. *Genetic Epidemiology* 2009; **33**(Suppl. 1):S1–S110.
13. Fang Y, Wang Y. Testing for genetic effect on functional traits by functional principal components analysis based on heritability. *Statistics in Medicine* 2009; **28**(29):3611–3625.
14. Eilers P, Marx B. Flexible smoothing with B-splines. *Statistical Science* 1996; **11**:89–121.
15. Chen H, Wang Y. A penalized spline approach to functional mixed effects model analysis. *Biometrics* 2010; DOI: 10.1111/j.1541-0420.2010.01524.x.
16. Chen Z, Dunson DB. Random effects selection in linear mixed models. *Biometrics* 2003; **59**:762–769.
17. Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* 1999; **86**:677–90.
18. Laird N, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
19. Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika* 2006; **93**:85–98.
20. Huang JZ, Liu L, Liu N. Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics* 2007; **16**(1):189–209.
21. Wu H, Zhang J. *Nonparametric Regression Methods for Longitudinal Data Analysis Mixed-effects Modeling Approaches*. Wiley: New York, 2006.

22. Louise T. Finding the observed information matrix when using the EM algorithm. *Journal of the American Statistical Association* 1982; **44**:226–233.
23. Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 2002; **11**(4):735–757.
24. Yu Y, Ruppert D. Penalized spline estimation for partially linear single index models. *Journal of the American Statistical Association* 2002; **97**:1042–1054.
25. Ruppert D, Wand M, Carroll R. *Semiparametric Regression*. Cambridge University Press: New York, 2003.
26. Krivobokova T, Kauermann G. A note on penalized splines with correlated Errors. *Journal of the American Statistical Association* 2007; **102**(480):1328–1337.
27. Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension* 2000; **36**(4):477–483.
28. Lasky-Su J, Lyon HN, Emilsson V, Heid IM, Molony C, Raby BA, Lazarus R, Klanderman B, Soto-Quiros ME, Avila L, Silverman EK, Thorleifsson G, Thorsteinsdottir U, Kronenberg F, Vollmert C, Illig T, Fox CS, Levy D, Laird N, Ding X, McQueen MB, Butler J, Ardlie K, Papoutsakis C, Dedoussis G, O'Donnell CJ, Wichmann HE, Celedón JC, Schadt E, Hirschhorn J, Weiss ST, Stefansson K, Lange C. On the replication of genetic associations: timing can be everything! *American Journal of Human Genetics* 2008; **82**:849–858.
29. Khoury M, Beaty H, Cohen B. *Fundamentals of Genetic Epidemiology*. Oxford University Press: New York, 1993.
30. Meyer K. Estimating covariance functions for longitudinal data using a random regression model. *Genetics Selection Evolution* 1998; **30**:221–240.
31. Rabe-Hesketh S, Skrondal A, Gjessing HK. Biometrical modelling of twin and family data using standard mixed model software. *Biometrics* 2008; **64**:280–288.
32. Shi G, Gu CC, Kraja AT, Arnett DK, Myers RH, Pankow JS, Hunt SC, Rao DC. Genetic effect on blood pressure is modulated by age the hypertension genetic epidemiology network study. *Hypertension* 2009; **53**(1):35–41.
33. Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* 2010; **66**:1069–1077.
34. Wand MP, Ormerod JT. On O'Sullivan penalised splines and semiparametric regression. *Australian and New Zealand Journal of Statistics* 2008; **50**:179–198.
35. Fan J, Huang T, Li R. Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* 2007; **102**(478):632–641.
36. Yao F, Lee TC. Penalizes spline models for functional principal component analysis. *Journal of the Royal Statistical Society*, *Series B* 2006; **68**:3–25.
37. James MG, Hastie TJ, Sugar AC. Principal component models for sparse functional data. *Biometrika* 2000; **87**(3):587–602.
38. Hall P, Müller HG, Wang JL. Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* 2006; **34**:1493–1517.