Biostatistics (2012), **13**, 3, *pp.* 482–496
doi:10.1093/biostatistics/kxr027
Advance Access publication on September 19, 2011

Semiparametric variance components models for genetic studies with longitudinal phenotypes

YUANJIA WANG*, CHIAHUI HUANG

Department of Biostatistics, Columbia University, New York, NY 10032, USA and Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
yuanjia.wang@columbia.edu

SUMMARY

In a family-based genetic study such as the Framingham Heart Study (FHS), longitudinal trait measurements are recorded on subjects collected from families. Observations on subjects from the same family are correlated due to shared genetic composition or environmental factors such as diet. The data have a 3-level structure with measurements nested in subjects and subjects nested in families. We propose a semiparametric variance components model to describe phenotype observed at a time point as the sum of a nonparametric population mean function, a nonparametric random quantitative trait locus (QTL) effect, a shared environmental effect, a residual random polygenic effect and measurement error. One feature of the model is that we do not assume a parametric functional form of the age-dependent QTL effect, and we use penalized spline-based method to fit the model. We obtain nonparametric estimation of the QTL heritability defined as the ratio of the QTL variance to the total phenotypic variance. We use simulation studies to investigate performance of the proposed methods and apply these methods to the FHS systolic blood pressure data to estimate age-specific QTL effect at 62cM on chromosome 17.

Keywords: Genome-wide linkage study; Multivariate longitudinal data; Penalized splines; Quantitative trait locus.

1. MOTIVATION

Multilevel or multivariate longitudinal data are commonly encountered in biological research. For example, in a family-based genetic study such as the Framingham Heart Study (FHS, Dawber *and others*, 1951), longitudinal trait measurements are recorded on subjects collected from families. Observations on subjects from the same family are correlated due to sharing genetic factors or environmental factors such as diet. The data structure has 3 levels with measurements nested in subjects and subjects nested in families. Some other examples of multivariate longitudinal data include clinical trials where multiple correlated variables such as messenger RNA expression levels and cluster of differentiation 4 (CD4) counts on a subject are collected over time (Lederman *and others*, 1998) or multiple informant data where information measuring the same underlying trait are collected from different sources (e.g. parents, teachers, and clinicians) over time (O'Brien and Fitzmaurice 2005). A common feature of these data is that there is an additional level of correlation to be accounted for: In the genetic studies, subjects within a family are correlated, and in multiple informant studies, different sources of information collected on the same subject are correlated.

*To whom correspondence should be addressed.

For univariate longitudinal data with independent subjects, there is a large body of literature on estimating population mean function or varying coefficients nonparametrically (see, e.g. Ruppert *and others*, 2003; Wu and Zhang, 2006). Recently, there is some attention on semiparametric modeling of multivariate longitudinal or functional data. Brumback and Rice (1998) used smoothing spline-based methods to analyze nested samples of functional data. Guo (2002) proposed functional mixed-effects model with functional random effects fitted by a Kalman filtering. Zhou *and others* (2008) modeled paired functional data by principal components using a reduced rank model. Baladandayuthapani *and others* (2008) and Staicu *and others* (2010) developed functional mixed-effects model-based Bayesian approaches for correlated multilevel spatial data. Di *and others* (2009) developed functional multivariate analysis of variance, which used a few functional principal components to reduce dimensionality. Aston *and others* (2010) proposed functional principal components analysis for linguistic pitch data. Greven *and others* (2010) proposed a computationally efficient functional principal components analysis applicable to functional data observed at multiple time points. However, all these approaches are not applicable to family studies where correlation of subjects in a family may depend on relationship between family members and genotypes at genetic markers.

In a family-based genetic study where longitudinal phenotype measurements are collected on subjects within families, usual genetic analyses proceed to analyze phenotype at each time point separately (Atwood *and others*, 2002) or adjust away the time trend and perform genetic analysis on the residuals (Levy *and others*, 2000). These approaches ignore potential gene–age interaction, which may contribute to inconsistencies in replications of genetic findings and loss of power (Lasky-Su *and others*, 2008; Shi and Rao, 2008).

One approach to analyze genetic linkage studies with longitudinal phenotypes is DeAndrade *and others* (2002), which incorporated repeated measures into a variance components model but treated age as nuisance parameters and accounted for the within-subject correlation across time points. However, this approach does not allow for direct modeling of age-specific genetic effect. A few existing work that do model time-specific genetic effect assumes a parametric function of the unknown effect over time (e.g. Shi and Rao, 2008; Zhang and Zhong, 2006). In some situations, a suitable parametric function is suggested by the biological underpinning of the development of a trait over time and offers meaningful interpretation of the biological structure. However, in many other situations, there may not be sufficient biological knowledge to warrant such a parametric form. Therefore, nonparametric methods on estimating quantitative trait locus (QTL) effect are useful. Approaches towards nonparametric modeling include Yang *and others* (2003), which averaged observations in several predivided intervals and used a regression spline approach. It is known that regression spline-based methods may be sensitive to number and location of knots. Zhao and Wu (2008) proposed a wavelet-based nonparametric approach to map QTL through a mixture model in structured population, which may not be easy to extend to general pedigrees.

In this work, we propose semiparametric variance components models for multilevel or multivariate longitudinal data. The proposed methods are suitable but not limited to longitudinal genetic linkage studies. In a linkage study, the QTL heritability is defined as the ratio of the genetic variance attributable to a QTL to the total phenotypic variance. To obtain nonparametric estimation of age-specific heritability, we treat the subject-specific processes as random nonparametric curves and estimate their covariance function nonparametrically. We use a kronecker product to specify the covariance function of a QTL effect to reflect genetic information between subjects in a family, while leaving the time component free of parametric assumptions. When there is no age-varying effect, the model reduces to a regular variance components model proposed for linkage data (Amos, 1994). We consider the population baseline function as an unspecified nonparametric function and use penalized splines (P-splines, O'Sullivan, 1986; Eilers and Marx, 1996) to estimate the nonparametric components of the model. Finally, we conduct simulation studies to investigate properties of the proposed methods and apply them to analyze the FHS longitudinal systolic blood pressure (SBP) data.

2. METHODS

2.1 *Traditional variance components model for univariate QTL data*

Genetic linkage analyses examine whether sharing more genetic material Identical by descent (IBD) at a locus leads to more similar phenotype values between 2 related individuals to infer whether the locus is near a gene affecting the phenotype. A traditional variance components model for a quantitative trait measured on subject j in family i is (Amos, 1994)

$$y_{ij} = \mu + \eta_{ij} + \delta_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \tag{2.1}$$

where μ is the population mean of the trait, η_{ij} is the major gene QTL effect, δ_{ij} is the additive polygenic effect, and ε_{ij} is the normal residual random measurement error with mean 0 and variance σ_ε^2. Information on the random QTL effect is modeled through the covariance between 2 related subjects in a pedigree. To be specific,

$$\text{Cov}(\eta_{ij}, \eta_{ij'}) = \pi^i_{jj'}\sigma^2_\eta, \quad \text{for } j, j' = 1, \dots, n_i, \quad \text{and} \quad \text{Cov}(\eta_{ij}, \eta_{i'j'}) = 0, \quad \text{for } i \neq i',$$

where $\pi^i_{jj'}$ denotes the proportion of alleles shared IBD between subjects j and j' in family i at a marker locus. The IBD-sharing matrix $\Pi_i = \{(\pi^i_{jj'}), j, j' = 1, \dots, n_i\}$ is computed based on marker genotypes and the recombination fraction between the marker and the putative QTL without using the phenotype data (Amos, 1994). In subsequent analyses, this matrix Π_i is regarded as can be estimated externally from marker genotypes before fitting a variance components model. The unknown variance of the QTL effect to be estimated from model (2.1) is σ^2_η.

The random residual polygenic effect δ_{ij} reflects residual genetic effect from other unlinked loci aside from the QTL. It is also characterized by its covariance between 2 subjects in a family, which is related to their relationship. To be specific, the covariance matrix of $\delta_i = (\delta_{i1}, \dots, \delta_{in_i})^T$ is specified through (Khoury *and others*, 1993)

$$\text{Cov}(\delta_{ij}, \delta_{ij'}) = 2K^i_{jj'}\sigma^2_\delta, \text{ for } j, j' = 1, \dots, n_i, \text{ and } \text{Cov}(\delta_{ij}, \delta_{i'j'}) = 0, \text{ for } i \neq i', \tag{2.2}$$

where $K^i_{jj'}$ is the known kinship coefficient between 2 relatives in a family. The kinship coefficient is defined as the probability of randomly drawing an allele from subject j, that is, IBD to an allele at the same locus randomly drawn from subject j' (Khoury *and others*, 1993). For example, the kinship coefficient is 1/4 for a full sibling pair and 1/8 for a half sibling pair. These coefficients are known given the relationship between relatives in a family. The unknown polygenic variance to be estimated is σ^2_δ. Since siblings have higher kinship coefficient than half siblings or cousins, covariance specification (2.2) suggests that the shared polygenic variance between a sibling pair is higher than that of a half sibling or cousin pair.

Parameters in model (2.1) are easily estimated through maximum likelihood. The QTL heritability is defined as the ratio of QTL variance to the total variance, that is, $h^2_\eta = \sigma^2_\eta/(\sigma^2_\eta + \sigma^2_\delta + \sigma^2_\varepsilon)$. Similarly, the residual polygenic heritability is defined as $h^2_\delta = \sigma^2_\delta/(\sigma^2_\eta + \sigma^2_\delta + \sigma^2_\varepsilon)$.

The traditional variance components model neither allows for age-dependent QTL effect nor accommodates longitudinal phenotypes. Several 2-step approaches have been proposed (Strauch *and others*, 2003) for longitudinal genetic linkage studies. In the first step, repeated measurements on the same subject are aggregated into several summary statistics such as subjects-specific intercepts and subject-specific slopes. In the second step, these summary statistics are used as outcomes in a conventional variance components model such as (2.1) to estimate genetic parameters including the QTL heritability. Such 2-step methods are less efficient than a joint estimation of longitudinal trend parameters and genetic parameters simultaneously in a single model. In addition, it is not always clear which summary statistic should be used, and these methods are subject to multiple comparisons when testing for QTL effect because one needs to test QTL effect on all summary statistics such as the intercept and the slope.

2.2 *A flexible semiparametric variance components model for longitudinal QTL data*

We now describe a semiparametric variance components model that can accommodate age-dependent QTL effect and repeated phenotype measurements. We start by introducing necessary notations to accommodate longitudinal phenotype. Let i index families, let j index subjects within a family, and let k index observations within a subject. Let n_i denote number of subjects in family i, let s_{ij} denote total number of measurements on the subject j in family i, and let $N_i = \sum_j s_{ij}$ denote number of observations from all subjects in family i. A semiparametric variance components model for longitudinal phenotype in a genetic study can be expressed as

$$y_{ijk} = \mu(t_{ijk}) + x_{ijk}^T\beta + w_{ijk}^T\alpha_i + \eta_{ij}(t_{ijk}) + u_{ijk}^T\delta_{ij} + \varepsilon_{ijk}, \tag{2.3}$$

where $\mu(t_{ijk})$ is a population baseline function, x_{ijk} is a vector of fixed effects for subject (i, j) such as gender and body mass index, $\alpha_i \sim N(0, D_\alpha)$ is a vector of random family-specific shared environmental effects such as diet (or common environmental effects, Khoury *and others*, 1993), w_{ijk} is a time-dependent design vector for α_i to accommodate potential age-varying common environmental effects, $\eta_{ij}(t_{ijk})$ is a random time-varying major QTL effect, δ_{ij} is a random residual polygenic effect, u_{ijk} is a time-dependent design vector for δ_{ij} to accommodate potential age-varying polygenic effect, and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ is a residual measurement error. We assume α_i, δ_{ij}, $\eta_{ij}(t_{ijk})$, and ε_{ijk} are independent.

The model (2.3) has 4 major components: a nonparametric baseline function $\mu(t)$, parametric fixed effects β, parametric random effects α_i and δ_{ij}, and nonparametric random effects $\eta_{ij}(t)$. Since our main interest lies in the QTL effect, to protect against model misspecification, it is modeled without imposing parametric assumptions on its functional form. For parsimony, the shared familial effect and residual polygenic effect are modeled parametrically through a low-dimensional design vector. For example, for a linear shared environmental effect, the design vector $w_{ijk} = (1, t_{ijk})^T$ and its time-dependent variance is $\sigma_\alpha^2(t_{ijk}) = w_{ijk}^T D_\alpha w_{ijk} = d_\alpha^{11} + 2d_\alpha^{12}t_{ijk} + d_\alpha^{22}t_{ijk}^2$, where d_α^{lm} is the (l, m)th component of D_α. Similarly, for a linear residual polygenic effect, the design vector $u_{ijk} = (1, t_{ijk})^T$ and its variance is $\sigma_\delta^2(t_{ijk}) = u_{ijk}^T D_\delta u_{ijk} = d_\delta^{11} + 2d_\delta^{12}t_{ijk} + d_\delta^{22}t_{ijk}^2$.

Using penalized splines, we model the QTL effect $\eta_{ij}(t)$ nonparametrically. Express $\eta_{ij}(t)$ in terms of linear combinations of a spline basis with random coefficients as

$$\eta_{ij}(t) = \Theta^T(t)\eta_{ij},$$

where $\Theta(t)$ is a q-dimensional vector of basis functions such as truncated polynomials and η_{ij} is the corresponding vector of random subject-specific coefficients. Extending the traditional variance components models for linkage studies (Amos, 1994) and random regression model for animal genetic studies (Meyer, 1998), the covariance matrix for η_{ij} is specified as

$$\begin{aligned}&\mathrm{Cov}(\eta_{ij}^{(l)}, \eta_{ij'}^{(l')}) = \pi_{jj'}^{i}\omega_{ll'}^2, \text{ for } j, j' = 1, \ldots, n_i, l, l' = 1, \ldots, q,\\ &\text{and } \mathrm{Cov}(\eta_{ij}^{(l)}, \eta_{i'j'}^{(l')}) = 0, \text{ for } i \neq i',\end{aligned} \tag{2.4}$$

where $\eta_{ij}^{(l)}$ is the lth component of the vector η_{ij} and $\pi_{jj'}^{i}$ is the (j, j')th element in the IBD-sharing matrix Π_i introduced in Section 2.1. The matrix $\Omega = \{(\omega_{ll'}^2), l, l' = 1, \ldots, q\}$ is the unknown covariance matrix of the QTL effect to be estimated from model (2.3).

The specification (2.4) implies that

$$\Psi_i \equiv \mathrm{Cov}(\eta_i, \eta_i^T) = \Pi_i \otimes \Omega, \tag{2.5}$$

where $\eta_i = (\eta_{i1}^T, \ldots, \eta_{in_i}^T)^T$. Therefore, one essentially models the covariance function of the QTL effect over time as a kronecker product of 2 sources: the subject-level source and the time-level source. The subject-level source is predicted by the IBD-sharing status at a marker locus based on the observed marker genotypes, that is, Π_i, while the time-level source involves the random basis coefficients and the unstructured QTL covariance matrix Ω. The QTL covariance between 2 subjects in a family is proportional to their IBD-sharing status. Such specification allows incorporation of genetic information as well as minimal assumptions on the functional form of the QTL variance as a function of time.

Note that by (2.5), we have $\mathrm{Cov}(\eta_{ij}(t), \eta_{ij'}(t)) = \pi^i_{jj'}\Theta(t)^T\Omega\Theta(t)$, which implies that at the same time point, the covariance of the QTL effect between subjects sharing more allele IBD at the locus is higher than the covariance between subjects sharing less allele IBD. In addition, if the covariance function does not change with time, that is, $\Theta(t) = 1$ and $\Omega = \omega^2$, which is a scalar value, then (2.3) reduces to the traditional variance components model (2.1) with a few more fixed and random effects. Using kronecker product to specify covariance matrix over time is also proposed for spatial–temporal data (Dutilleul and Pinel-Alloul, 1996) and multiple informant data (O'Brien and Fitzmaurice, 2005). It is well known that accurate modeling of covariance process improves estimation efficiency of the population baseline function. We will also show this effect through simulation studies in Section 3.

The age-dependent random residual polygenic effect δ_{ij} is modeled parametrically through a time-dependent design vector u_{ijk}. Similar to the time-invariant case, the covariance of δ_{ij} is specified as

$$\mathrm{Cov}(\delta_{ij}, \delta_{ij'}) = 2K^i_{jj'}D_\delta, \text{ for } j, j' = 1, \ldots, n_i, \text{ and } \mathrm{Cov}(\delta^{(l)}_{ij}, \delta^{(l')}_{i'j'}) = 0, \text{ for } i \neq i',$$

where $K^i_{jj'}$ is again the known kinship coefficient between 2 relatives in a family introduced in Section 2.1. The unknown polygenic covariance to be estimated is D_δ.

Using a linear combination of spline basis with fixed coefficients, we express the population mean function as $\mu(t) = \Theta^T(t)\mu$. To facilitate further computation, we write (2.3) in a matrix form as

$$Y_i = X_i\beta + B_i\mu + W_i\alpha_i + U_i\delta_i + Z_i\eta_i + \varepsilon_i,$$

$$\alpha_i \sim N(0, D_\alpha), \delta_i \sim N(0, 2K_i \otimes D_\delta), \eta_i \sim N(0, \Pi_i \otimes \Omega), \varepsilon_i \sim N(0, V_i), \tag{2.6}$$

where $Y_i = (Y_{ijk})^T_{j=1,\ldots,n_i,k=1,\ldots,s_{ij}}$ is a vector of phenotype measurements for all subjects in the family i at all occasions, X_i is a design matrix for fixed effects, $B_i = (\Theta(t_{ijk}))^T_{j=1,\ldots,n_i,k=1,\ldots,s_{ij}}$ is a matrix of basis functions of the population baseline, U_i and W_i are stacked matrices of u_{ijk} and w_{ijk}, $Z_i = \mathrm{diag}\{\zeta(t_{i1}), \ldots, \zeta(t_{in_i})\}$, where $\zeta(t_{ij}) = (\Theta(t_{ij1}), \ldots, \Theta(t_{ijs_{ij}}))^T$, is a $N_i \times n_i k$ block diagonal matrix of basis functions for the QTL effect, $K_i = (K^i_{jj'})_{j,j'=1,\ldots,n_i}$ is the matrix of kinship coefficients for all subjects in family i, and $V_i = \sigma_\varepsilon^2 I_{N_i}$. Since the covariance matrix for the QTL effect Ω is unstructured, the number of parameters involved can be large. For example, with a q-dimensional basis, there are $q(q+1)/2$ distinct parameters involved in Ω.

From model (2.3), the age-specific QTL variance is

$$\sigma^2_\eta(t) = \Theta^T(t)\Omega\Theta(t),$$

and the age-specific QTL heritability is

$$h^2_\eta(t) = \frac{\sigma^2_\eta(t)}{\sigma^2_T(t)} = \frac{\Theta^T(t)\Omega\Theta(t)}{\sigma^2_T(t)}, \tag{2.7}$$

where the total phenotypic variance is $\sigma_T^2(t) = \sigma_\alpha^2(t) + \sigma_\delta^2(t) + \sigma_\eta^2(t) + \sigma_\varepsilon^2$. To see flexibility in estimating QTL variance offered by an unstructured covariance matrix $\mathbf{\Omega}$, consider an example using a linear truncated polynomial basis with knots τ_k, that is, $\Theta(t) = (1, t, (t-\tau_1)_+, \dots, (t-\tau_K)_+)^T$. If $\mathbf{\Omega}$ is restricted as a diagonal matrix, that is, $\mathbf{\Omega} = \text{diag}(\omega_0^2, \dots, \omega_{K+1}^2)$, then

$$\sigma_\eta^2(t) = \omega_0^2 + \omega_1^2 t^2 + \omega_2^2 (t-\tau_1)_+^2 + \ldots + \omega_{K+1}^2 (t-\tau_K)_+^2.$$

Therefore, the QTL variance is restricted to be nondecreasing, which may not be desirable. Nonetheless, if $\mathbf{\Omega}$ is completely unrestricted without a roughness penalty, the fitted subject-specific curves and QTL variance may be highly nonsmooth.

2.3 *Penalized spline estimation*

Penalized spline-based methods express an arbitrary function as a linear combination of high-dimensional basis functions and penalize some measure of roughness of the fitted curves to provide smooth estimation. To be specific, given the variance components, we estimate the population baseline function by minimizing the penalized weighted least squares,

$$\sum_i \left[(Y_i - \tilde{X}_i \tilde{\mu})^T \tilde{V}_i^{-1} (Y_i - \tilde{X}_i \tilde{\mu}) + \log|\tilde{V}_i| \right] + \lambda_1 \tilde{\mu}^T \Delta_1 \tilde{\mu},$$

where $\tilde{X}_i = (X_i, B_i)$, $\tilde{\mu} = (\beta^T, \mu^T)^T$, $\tilde{V}_i = W_i D_\alpha W_i^T + 2U_i K_i \otimes D_\delta U_i^T + Z_i \Psi_i Z_i^T + V_i$, to obtain

$$\widehat{\tilde{\mu}} = \sum_i (\tilde{X}_i^T \tilde{V}_i^{-1} \tilde{X}_i + \lambda_1 \Delta_1)^{-1} \tilde{X}_i^T \tilde{V}_i^{-1} Y_i.$$

Here, λ_1 is a smoothing parameter and Δ_1 is a penalty matrix depending on the chosen basis for the mean function. The smoothing parameter controls the amount of smoothing, where a large value encourages fitting a polynomial curve and a small value encourages interpolation therefore a wiggly curve. For the pth-order truncated polynomial basis with $(p+1)$th-order penalty, the penalty matrix is a diagonal matrix with the first $p+1$ diagonal elements zero and the remaining diagonal elements one. For B-spline basis with dth-order penalty, the penalty matrix is a dth-order difference matrix as defined in Eilers and Marx (1996).

To estimate the variance components parameters, we use the Expectation Maximization algorithm. Regard random effects α_i, δ_i, and η_i as missing data and consider the following penalized joint likelihood (see, e.g. Wu and Zhang, 2006):

$$\sum_i \left[e_i^T V_i^{-1} e_i + \delta_i^T (2K_i \otimes D_\delta)^{-1} \delta_i + \eta_i^T \Psi_i^{-1} \eta_i + \log|K_i \otimes D_\delta| + \log|V_i| + \log|\Psi_i| \right]$$
$$+ \lambda_2 \sum_i \eta_i^T \Delta_2 \eta_i, \qquad (2.8)$$

where $e_i = (Y_i - X_i\beta - B_i\mu - W_i\alpha_i - U_i\delta_i - Z_i\eta_i)$, λ_2 is a smoothing parameter for the subject-specific QTL curves, and Δ_2 is a penalty matrix related to the chosen basis. The smoothing parameter here controls smoothness of the fitted subject-specific QTL curves and the covariance function. It is motivated by the assumption that the random major gene QTL effects are realizations of a Gaussian process with a smooth covariance function. Here, with a pth-order truncated polynomial basis and K knots, the penalty matrix of the subject-specific curves is $\Delta_2 = \text{diag}(\mathbf{1}_{n_i}) \otimes \text{diag}(\mathbf{0}_{p+1}, \mathbf{1}_K)$.

We describe details of the EM algorithm and selection of smoothing parameters in the Supplementary Appendix available at *Biostatistics* online. After convergence, the fitted age-specific QTL variance is

$$\hat{\sigma}_{\eta}^{2}(t) = \Theta^{T}(t)\widehat{\Omega}\Theta(t),$$

and the age-specific QTL heritability is

$$\hat{h}_{\eta}^{2}(t) = \frac{\hat{\sigma}_{\eta}^{2}(t)}{\hat{\sigma}_{T}^{2}(t)} = \frac{\Theta^{T}(t)\widehat{\Omega}\Theta(t)}{\hat{\sigma}_{T}^{2}(t)}, \tag{2.9}$$

where the total phenotypic variance is $\hat{\sigma}_{T}^{2}(t) = \hat{\sigma}_{\alpha}^{2}(t) + \hat{\sigma}_{\delta}^{2}(t) + \hat{\sigma}_{\eta}^{2}(t) + \hat{\sigma}_{\varepsilon}^{2}$.

3. Simulations

In this section, we present simulation studies to evaluate performance of the proposed methods. The family structure, sample size, and observed assessment schedules were taken as the same as observed in the FHS data and are described in the next section. We simulated trait data from model (2.3), where we used the estimated population mean function from the FHS (see Figure 1) as $\mu(t)$. We used the observed genotypes at the marker GATA25A04 on chromosome 17 to simulate the QTL effect. Specifically, η_{ij} was simulated from a multivariate normal distribution with covariance

$$\mathrm{Cov}(\eta_{ijk}, \eta_{ij'k'}) = \pi_{jj'}^{i} \nu(t_{ijk})\nu(t_{ij'k'}),$$

where $\nu(t) = b_1\sqrt{\exp[a_1^2(t-a_2)^2]}$ and $\pi_{jj'}^{i}$ is the proportion of alleles shared IBD between subjects j and j' in family i computed from the observed genotypes at the marker GATA25A04 (chromosome 17) from the FHS subjects. We first examined 3 scenarios of the time-varying QTL effect with $a_1 = (-0.08, -0.03, -0.02)^T$, $a_2 = (45, 45, 45)^T$, and $b_1 = (4, 1.5, 0.67)^T$, and the maximum QTL heritability over time was 0.94, 0.64, and 0.31 in these examples. In the next 3 examples, we simulated a polygenic effect with standard deviation (SD) 1.5. The parameters for $\nu(t)$ were $a_1 = (-0.08, -0.03, -0.02)^T$, $a_2 = (45, 45, 45)^T$, and $b_1 = (5, 2.2, 1)^T$. The maximum heritability over time was 0.88, 0.59, and 0.23, respectively. We depict the true age-varying heritability for each of these examples in a Supplementary Figure available at *Biostatistics* online. The residual variance was one in all examples.

To examine performance of the estimated mean function and the QTL variance function in each example, we computed the average mean squared error (AMSE) of the estimated functions across the 200 simulations of the mean squared error defined as $\mathrm{MSE}(f) = \frac{1}{\sum_{ij} s_{ij}} \sum_{ijk}[\hat{f}(t_{ijk}) - f(t_{ijk})]^2$. Table 1 summarizes the AMSEs of the population baseline function and the QTL variance function. We can see that while the AMSE for the mean function was small for all examples, it increased when the true QTL variance increases due to larger total variance of the outcomes. We examined the pointwise coverage probability of the confidence interval (CI) of $\hat{\mu}(t)$ at age 45 and found that the empirical coverage adheres to the nominal level (Table 1). The AMSEs of the QTL variance also increased when the true QTL variance is larger due to greater variability in the outcome. We compared the efficiency of estimating population mean function using the proposed covariance structure with using a working independent covariance. From Table 1, we see that the efficiency improvement in AMSE($\hat{\mu}(t)$) ranges from 10% to 59% in the 6 examples.

Next, we present a sensitivity analysis to examine effect of violating normality assumption of the residuals. We generated ε_{ijk} in (2.6) from a t distribution with degrees of freedom 20 and other simulation parameters were the same as in scenario 3. The results show that the AMSE of the mean and QTL variance function increased slightly with a misspecified residual distribution. The AMSE of the mean function

Table 1. *Properties of the estimated functions*

	Max	Max	AMSE	AMSE	Eff	AMSE	50%	
Ex.	$\sigma_{\eta}^{2}(t)$†	$h_{\eta}^{2}(t)$‡	$\hat{\mu}(t)$	$\hat{\mu}_{ind}(t)$§	improv¶	$\hat{\sigma}_{\eta}^{2}(t)$	coverage‖	$\hat{\sigma}^{2}$
1	16.00	0.94	0.021	0.051	59%	0.935	0.47	0.954
2	2.25	0.64	0.014	0.017	19%	0.036	0.46	0.929
3	0.45	0.31	0.008	0.009	10%	0.020	0.48	0.930
4	25.00	0.88	0.057	0.017	47%	2.548	0.45	0.989
5	2.84	0.59	0.037	0.055	32%	0.411	0.42	0.948
6	1.00	0.23	0.024	0.034	29%	0.304	0.51	0.930

†Maximum QTL variance over the range of t, that is, $\max_t \sigma_{\eta}^{2}(t)$.

‡Maximum QTL heritability over the range of t, that is, $\max_t h_{\eta}^{2}(t)$.

§$\hat{\mu}(t)$ estimated by assuming a working independent covariance matrix.

¶Efficiency improvement measured by [AMSE{$\hat{\mu}_{ind}(t)$}-AMSE{$\hat{\mu}(t)$}]/AMSE{$\hat{\mu}_{ind}(t)$}.

‖Coverage probability of 50% CI of $\hat{\mu}(t)$ at age 45.

increased from 0.009 in example 3 to 0.011 and the AMSE of the QTL variance function increased from 0.020 to 0.029. When we generated the random residuals from a t distribution with 4 degrees of freedom, the AMSE of the mean function increased to 0.017 and the AMSE of the QTL variance function increased to 0.09, indicating some efficiency loss when the error distribution has a heavy tail.

4. Analysis of the FHS data

We applied proposed methods to analyze the FHS SBP data. The FHS (Dawber *and others*, 1951) is an ongoing prospective study of risk factors for cardiovascular disease (CVD), which started in 1948. In the FHS, healthy subjects were recruited and followed over a long period of time. The objective of the FHS is to identify common risk factors or characteristics that contribute to CVD in a large population-based sample. High blood pressure is considered as a major risk factor for stroke and heart disease, and it affects about one-third of the US adult population. SBP is a complex trait that may have complicated etiology. Previous literature suggests a substantial genetic contribution to SBP (Levy *and others*, 2000).

In the late 1990s, genome-wide linkage studies were conducted in the FHS family samples. Levy *and others* (2000) identified a locus on chromosome 17 (GATA25A04, 62cM) with a high logarithm (base 10) of odds score. Here, we fit the model (2.3) to the longitudinal SBP data and estimate the age-specific QTL effect at this locus. We computed the IBD-sharing matrix Π_i using SOLAR (Almasy and Blangero, 1998) and the kinship coefficient matrix using R package "Kinship" (Atkinson and Therneau, 2009).

We analyzed observations between age 30 and 60. There were 318 subjects from 105 families with 1559 observations. On average, there were 3.0 subjects in each family and 4.9 measurements on each subject. The size of families ranges from 2 to 9. The age of the participants at the first visit ranges from 30 to 51 and the mean age for all subjects at all visits was 45.85 years. The mean observed SBP was 121.31 mm Hg. We show a scatter plot of SBP versus age for 100 randomly selected subjects in the left panel of Figure 1.

Before fitting a semiparametric variance components model, we conducted several exploratory analyses. First, we estimated the ratio of genetic variance to the total variance (heritability) by a polynomial model in 3 types of relatives: (1) first-degree relatives with kinship coefficients 1/4 such as a parent–offspring pair or a full sibling pair; (2) second-degree relatives with kinship coefficient 1/8 such as half sibling pairs or aunt/uncle–niece/newphew pairs; and (3) other more distant relatives with kinship coefficient 1/16 such as first cousins. As expected, from the upper left panel in Figure 2, we see that

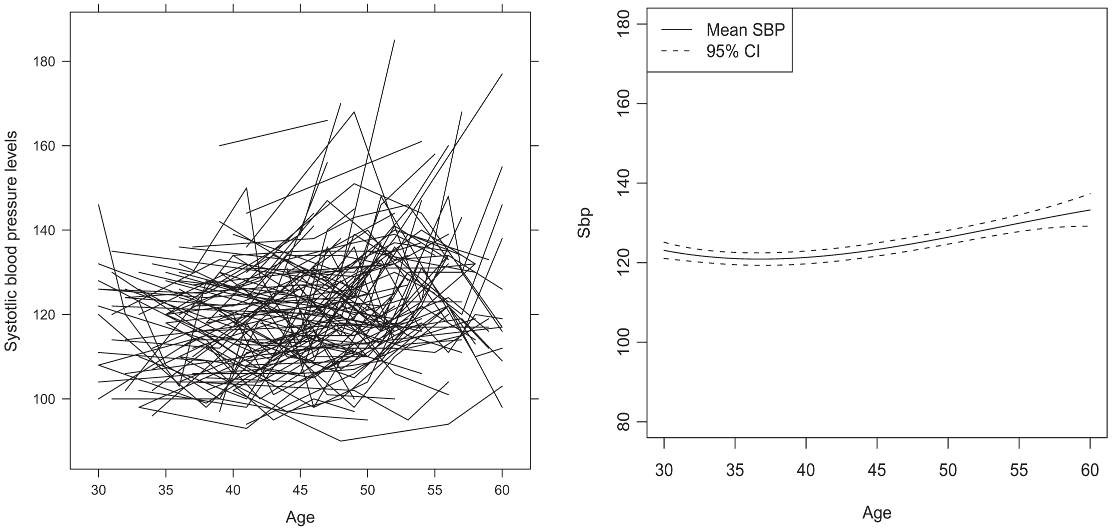

Fig. 1. Scatterplot of the SBP versus age for 100 randomly selected subjects in the FHS (left panel) and estimated age-specific baseline SBP in the FHS (right panel).

the heritability in the first-degree relatives is highest followed by second-degree relatives and other relatives. The average estimated heritability over time is 0.31, 0.27, and 0.07 in first-degree, second-degree, and other relatives, respectively. Second, we compute a Pearson correlation of SBP between relative pairs in the 3 groups of relatives. Figure 2 displays the scatter plot of age-adjusted SBP between relative pairs in each group. The correlation was 0.19, 0.11, and 0.02 in first-degree, second-degree, and other relatives, respectively. The correlation of adjusted SBP between second-degree relative pairs is about a half of that in the first-degree pairs. From these exploratory analyses, we observe larger heritability and correlation in closely related relatives comparing to more distant relatives, which is consistent with the model (2.6).

To fit a semiparametric variance components model, we used cubic truncated polynomial basis with 10 knots for both $\mu(t)$ and $\eta_i(t)$. We present the population baseline function of SBP in the right panel of Figure 1. The mean SBP increased from 123.10 mm Hg at age 30 to 133.25 mm Hg at age 60. We assumed a linear shared environmental effect and polygenic effect. The SD of the residual measurement error was estimated to be 8.76. We present the age-specific QTL variance at marker GATA25A04, and its CI computed by bootstrap in the upper left panel of Figure 3. The QTL variance ranges from 41.1 to 259.8. It increases slowly from age 30 to 50 and after 50, the rate of increase picks up. We computed the age-specific QTL heritability by (2.9) and present the estimates in the upper right panel of Figure 3. The heritability was 0.13 (95% CI: 0.08–0.18) at age 30 and increased to 0.34 (95% CI: 0.24–0.44) at age 60. The difference of heritability at age 60 and 30 was 0.21 with a 95% CI of (0.12–0.31), suggesting a significant increase from age 30 to 60. The long-term overall heritability of SBP was estimated to be 0.57 (95% CI: 0.53–0.61) in Levy *and others* (2000). There is no attempt to estimate age-specific heritability at the marker GATA25A04 in the literature. Here, the average QTL heritability over time at this marker was estimated to be 0.22 (95% CI: 0.18–0.27), which suggests potential genetic contribution to SBP at other unlinked loci in addition to GATA25A04 since this marker does not fully explain the variance of the genetic component.

We provide 2 views of the 3D surface plot of the correlation function of the QTL effect in 2 lower panels in Figure 3. We also show a 2D graph of the correlation versus time lag at different ages in the supplementary material available at *Biostatistics* online. The autocorrelation decreases when the lag between

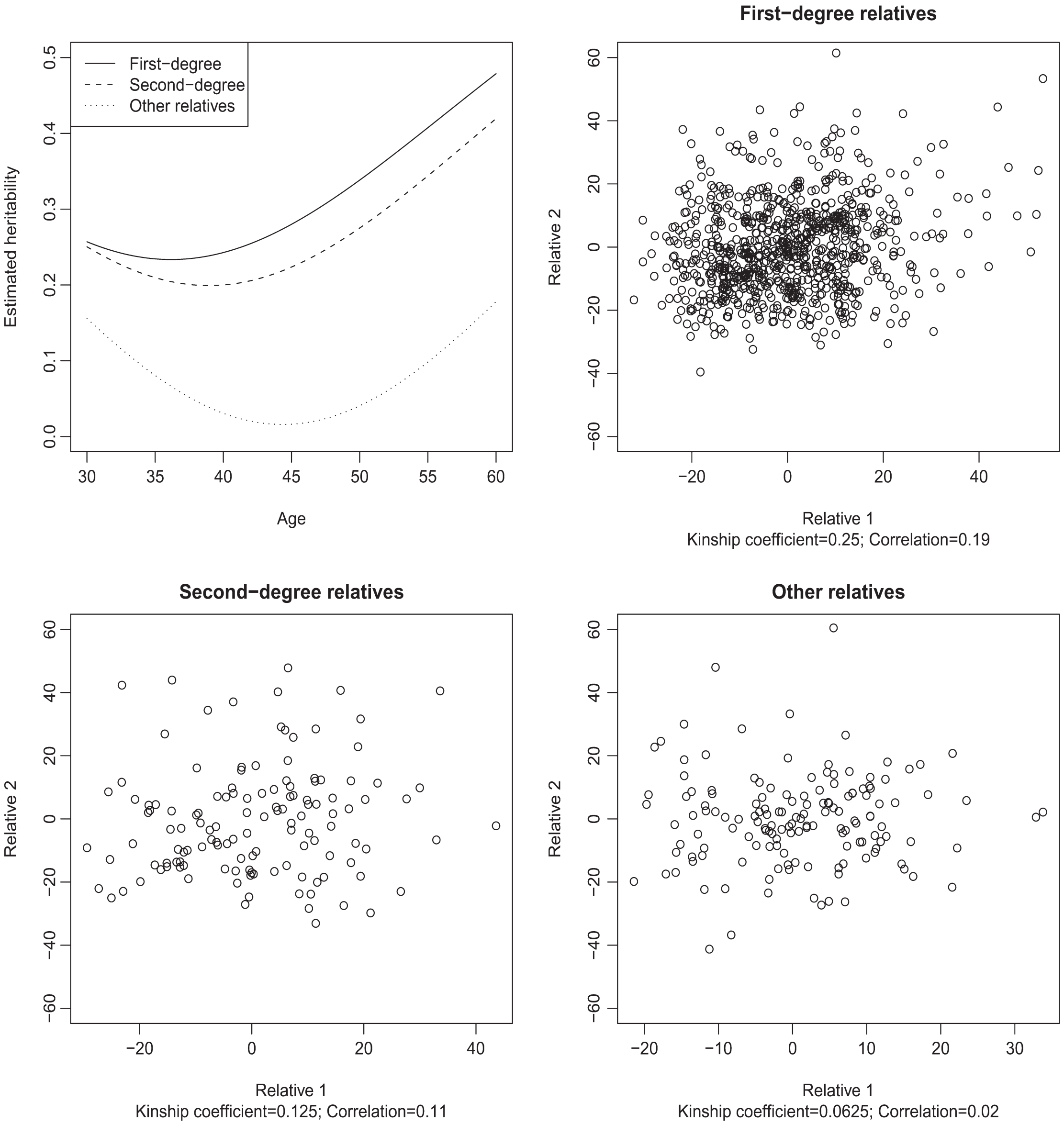

Fig. 2. Estimated ratio of genetic variance to total variance (heritability) in 3 groups of relatives (upper left) and scatterplots of relative pairs in 3 groups of relatives (upper right, lower left, and lower right).

2 time points increases. It is seen from the 2D plot that depending on the starting age, the autocorrelation function decreases at different rate. Therefore, the QTL process does not have a simple structure such as First order autoregressive model. This analysis illustrates the advantage of allowing a flexible nonstructured covariance matrix Ω in (2.5).

Next, we demonstrate the benefit of using a mixed-effects model to predict individual phenotype trajectories from different sources of information. We first used the data from all siblings to fit the model

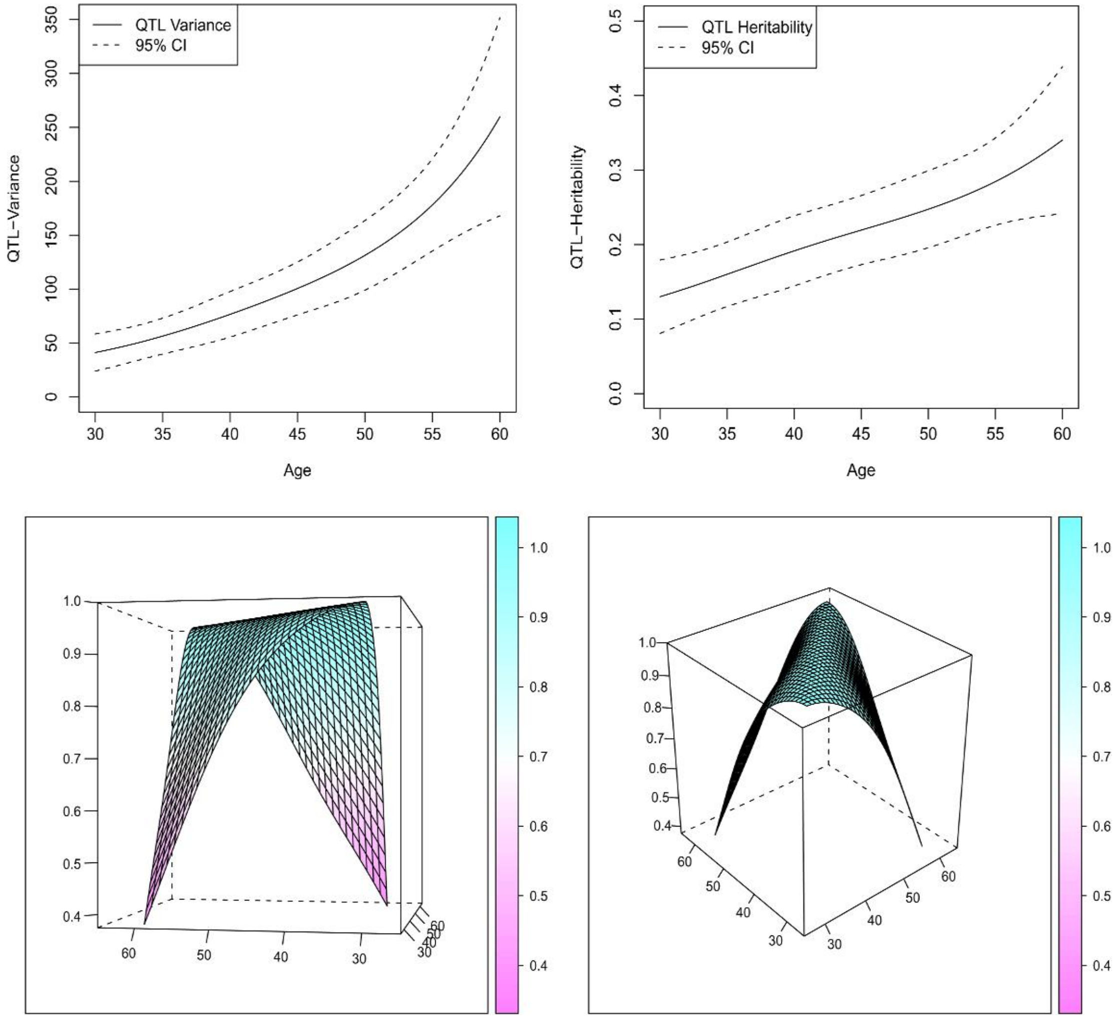

Fig. 3. Estimated age-specific QTL variance (upper left), age-specific heritability (upper right) at 62cM on chromosome 17 (marker GATA25A04) in the FHS, and 2 views of the 3D surface plot of the autocorrelation function of the QTL effect at this marker (lower left and lower right).

and predict individual trajectories. We computed a root mean prediction error as $\sqrt{\sum_{ijk}(y_{ijk} - \hat{y}_{ijk})^2/N}$, where $\hat{y}_{jik}$ is the predicted phenotype for a subject and compared it with the root mean prediction error obtained using data on all cousins. The root mean prediction error was 7.69 using sibling data and increased to 8.40 when using cousins, which indicated a slight gain in prediction accuracy using more closely related relatives. We present an individual's predicted trajectory obtained from different sources of data in Figure 4: the trajectory predicted from siblings and cousins as well as the population mean curve. It can be seen that the trajectory predicted from the siblings is closer to the observed values than the one predicted from the cousins.

Although here we focused on estimating age-specific QTL effect at a locus, we also performed a likelihood ratio test of the QTL effect, that is, $H_0 : \Omega = 0$ versus $H_a : \Omega > 0$. The test statistic is

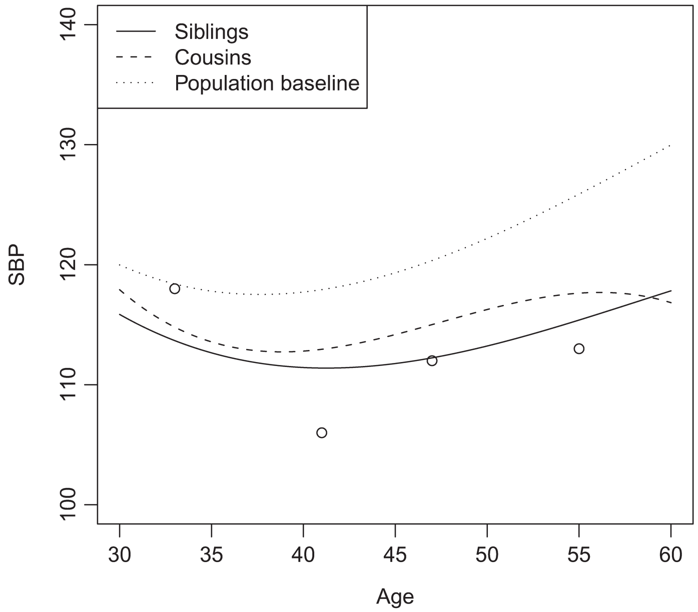

Fig. 4. Predicted SBP from siblings and cousins.

$T = 2\sup L_{H_a} - 2\sup L_{H_0}$, where L_{H_0} and L_{H_a} are the log likelihood under the null and the alternative hypothesis, respectively. Crainiceanu and Ruppert (2004) and Crainiceanu *and others* (2005) showed that the distribution of the likelihood ratio statistic for penalized spline regression is nonstandard due to a lack of independence under the null and suggested a simulation-based approach for a linear mixed-effects model with one variance component. Here, we used a parametric bootstrap procedure to compute the $p$value. We simulated the data under the null distribution by assuming $\Omega = 0$ in model (2.3). To be specific, we simulated the bth copy of the bootstrap sample under the null as

$$Y_{ijk}^{(b)} = \hat{\mu}(t_{ijk}) + w_{ijk}^T \alpha_i^{(b)} + u_{ijk}^T \delta_{ij}^{(b)} + \varepsilon_{ijk}^{(b)},$$

where $\alpha_i^{(b)} \sim N(0, \widehat{D}_\alpha)$, $\delta_i^{(b)} \sim N(0, 2K_i \otimes \widehat{D}_\delta)$, and $\varepsilon_{ijk}^{(b)} \sim N(0, \hat{\sigma}_\varepsilon^2)$. The mean function and the variance components such as $\widehat{D}_\alpha$ and $\widehat{D}_\delta$ were estimated from the more general model (2.3) allowing for a QTL effect. We then computed the likelihood ratio statistic using the bootstrap samples and computed the p value as the sample proportion of the bootstrap test statistics greater than the observed test statistic. We simulated 200 bootstrap samples and the p value was less than 0.05, suggesting a significant QTL effect.

5. Discussion

In this work, we propose flexible variance components models for genetic studies with longitudinal phenotypes. The population mean function and QTL heritability are estimated nonparametrically. We incorporate genetic information by using IBD-sharing matrix at genetic markers and kinship coefficient matrix, and we leave the age component of the QTL effect as an unrestricted function. The proposed methods contribute to more flexible modeling of age-specific genetic effect using variance components models. Although we present methods for longitudinal phenotypes, it is straightforward to apply these methods to cross-sectional phenotypes.

The computational speed of the proposed methods mainly depends on the size of families, number of knots, and the convergence rate of the EM algorithm. The dimension of the family-specific covariance matrix increases with the family size; therefore, the algorithm may be slow for large pedigrees. For the FHS data analyzed here (family size ranges from 2 to 9 and a total sample size of 1559), the EM algorithm usually converges in about 30 iterations. For large pedigrees, the sparse representation of a kinship coefficient matrix in the "Kinship" package (Atkinson and Therneau, 2009) can be used to improve computational efficiency. In addition, singular value decomposition may be used to compute inverse of a large matrix and a combination of EM algorithm and Newton–Raphson iteration may further improve computational efficiency.

Here, we treated the residual polygenic effect as a linear function of age, which implies that the residual genetic effect from other unlinked loci is linear. One can fit a nonparametric time-varying polygenic effect by handling it similarly as the QTL effect. That is, to express the polygenic effect as a linear combination of basis functions with random coefficients (Wang, 2011). The covariance matrix of the coefficients will be constructed by the kronecker product of a kinship coefficient matrix and an unknown polygenic covariance. An adapted EM algorithm can be used for the estimation. However, a disadvantage of such an approach is the increase in computational burden.

The proposed methods are illustrated using genetic linkage data. These approaches can also be applied to estimate covariance function for other multilevel longitudinal or functional data such as spatial data and multiple informant data where the covariance function is specified as a kronecker product of 2 sources. It is also easy to extend model (2.3) to account for fixed effects with varying coefficients by penalized splines (see e.g. Chen and Wang, 2011).

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The data were obtained from the FHS 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). The authors wish to thank the associate editor and 2 anonymous reviewers for their insightful comments, which significantly contributed to improving the quality of this work. *Conflict of Interest:* None declared.

FUNDING

National Institutes of Health (AG031113-01A2 and NS073670-01 to Y.W.).

REFERENCES

ALMASY, L. AND BLANGERO, J. (1998). *American Journal of Human Genetics* **62**, 1198–1211.

AMOS, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* **54**, 535–543.

ASTON, J. A. D., CHIOU, J.-M. AND EVANS, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society Series C* **59**, 297–317.

ATKINSON, B. AND THERNEAU, T. (2009). *kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package version 1.1.0-23*. http://CRAN.R-project.org/package=kinship.

Atwood, L., Heard-Costa, N., Cupples, L., Jaquish, C., Wilson, P. and D'Agostine, R. (2002). Genome-wide linkage analysis of body mass index across 28 years of the Framingham Heart Study. *American Journal of Human Genetics* **71**, 1044–1050.

Baladandayuthapani, V., Mallick, B. K., Hong, M. Y., Lupton, J. R., Turner, N. D. and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* **64**, 64–73.

Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association* **93**, 961–994.

Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics* (in press). PMC2948587. doi: 10.1111/j.1541-0420.2010.01524.x

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **65**, 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G. and Wand, P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103.

Dawber, T. R., Meadors, G. F. and Moore, F. E. J. (1951). Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health and the Nation's Health* **41**, 279–286.

DeAndrade, M., Guéguen, G., Visvikis, S., Sass, S., Siest, G. and Amos, C. I. (2002). Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. *Genetic Epidemiology* **22**, 221–232.

Di, C., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics* **3**, 458–488.

Dutilleul, P. and Pinel-Alloul, B. (1996). A doubly multivariate model for statistical analysis of spatio-temporal environmental data. *Environmetrics* **7**, 551–566.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines. *Statistical Science* **11**, 89–121.

Greven, G., Crainiceanu, C., Caffo, B. and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics* **4**, 1022–1054.

Guo, W. (2002). Functional mixed effects models. *Biometrics* **58**, 121–128.

Khoury, M., Beaty, H. and Cohen, B. (1993). *Fundamentals of Genetic Epidemiology*. New York: Oxford University Press.

Lasky-Su, J., Lyon, H. N., Emilsson, V. *and others* (2008). On the replication of genetic associations: timing can be everything!, *American Journal of Human Genetics* **82**, 849–858.

Lederman, M. M., Connick, E. *and others* (1998). Immunological responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine and ritonavir: results of AIDS Clinical Trials Group Protocol 315. *Journal of Infectious Diseases* **178**, 70–79.

Levy, D., DeStefano, A. L., Larson, M. G., O'Donnell, C. J., Lifton, R. P., Gavras, H., Cupples, L. A. and Myers, R. H. (2000). 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* **36**, 477–483.

Meyer, K. (1998). Estimating covariance functions for longitudinal data using a random regression model. *Genetics Selection Evolution* **30**, 221–240.

O'Brien, L. M. and Fitzmaurice, G. M. (2005). Regression models for the analysis of longitudinal Gaussian data from multiple sources. *Statistics in Medicine* **24**, 1725–1744.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502–518.

RUPPERT, D., WAND, M. P. AND CARROLL, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

SHI, G. AND RAO, D. C. (2008). Ignoring temporal trends in genetic effects substantially reduces power of quantitative trait linkage analysis. *Genetic Epidemiology* **32**, 61–72.

STAICU, A., CRAINICEANU, C. AND CARROLL, C. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics* **11**, 177–194.

STRAUCH, J., GOLLA, A., WILCOX, M. A. AND BAUR, M. P. (2003). Genetic analysis of phenotypes derived from longitudinal data: presentation group 1 of Genetic Analysis Workshop 13. *Genet Epidemiol* **25** (suppl 1), S5–S17.

WANG, Y. (2011). Flexible estimation of covariance function by penalized spline with application to longitudinal family data. *Statistics in Medicine* (in press). doi:10.1002/sim.4236.

WU, H. AND ZHANG, J. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis Mixed-Effects Modeling Approaches*. New York: Wiley.

YANG, Q., CHAZARO, I., CUI, J., GUO, C. Y., DEMISSIE, S., LARSON, M., ATWOOD, D. L., CUPPLES, L. A. AND DESTEFANO, A. L. (2003). Genetic analysis of longitudinal phenotype data: a comparison of univariate methods and a multivariate approach. *BMC Genetics* **4** (Suppl 1), S29.

ZHANG, H. AND ZHONG, X. (2006). Linkage analysis of longitudinal data and design consideration, *BMC Genetics* **7**, 37.

ZHAO, W. AND WU, R. (2008). Wavelet-based nonparametric functional mapping of longitudinal curves. *Journal of the American Statistical Association* **103**, 714–725.

ZHOU, L., HUANG, J. AND CARROLL, R. (2008). Joint modeling of paired sparse functional data using principal components. *Biometrika* **95**, 601–619.

[*Received January 17, 2011; revised July 24, 2011; accepted for publication July 26, 2011*]