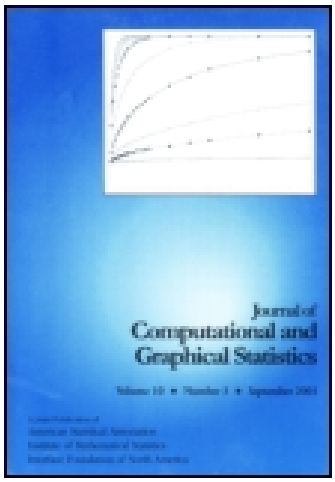


Journal of Computational and Graphical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

Functional Data Model for Genetically Related Individuals with Application to Cow Growth

Edwin Lei Ph.D. student[a], Fang Yao Professor[b], Nancy Heckman Professor[c] & Karin Meyer Principal Scientist[d]

[a] Department of Statistical Sciences, University of Toronto, ON, M5S 3G3, Canada

[b] Department of Statistical Sciences, University of Toronto, ON, M5S 3G3, Canada,

[c] Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

[d] Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia.

Accepted author version posted online: 08 Aug 2014.

To cite this article: Edwin Lei Ph.D. student, Fang Yao Professor, Nancy Heckman Professor & Karin Meyer Principal Scientist (2014): Functional Data Model for Genetically Related Individuals with Application to Cow Growth, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2014.948180

To link to this article: http://dx.doi.org/10.1080/10618600.2014.948180

Functional Data Model for Genetically Related Individuals with Application to Cow Growth

Edwin LEI, Fang YAO, Nancy HECKMAN, Karin MEYER

August 5, 2014

We propose a new version of functional data model for analyzing familial related individuals, where the within-subject correlation depends smoothly on a covariate such as age and the between-subject correlation follows family-wise genetic association. Our motivating example concerns measurements of weight as a function of age in sibling cows from independent families. Observations are sparsely sampled from trajectories of a phenotype contaminated with measurement error, where the phenotypic trajectory consists of a genetic component and an environmental component. By combining information across individuals, the genetic and environmental covariance are estimated via smoothing techniques. We study the genetic and environmental effects using principal component analysis, taking into account the genetic correlation to enhance the subject-level signal extraction. We show via the real data and simulations that incorporating the correlation structure improves predictions of individual phenotypic trajectories.

Key words: Functional principal components, Genetic relationship, Smoothing, Sparse functional data.

Edwin Lei is Ph.D. student, Department of Statistical Sciences, University of Toronto, ON, M5S 3G3, Canada; Fang Yao is Professor, Department of Statistical Sciences, University of Toronto, ON, M5S 3G3, Canada (e-mail: fyao@utstat.toronto.edu); Nancy Heckman is Professor, Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; and Karin Meyer is Principal Scientist, Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia.

1. INTRODUCTION

Functional data analysis (FDA) has attracted substantial research interest and has provided powerful tools to study data arising from a collection of curves rather than from scalars or vectors. Ramsay and Silverman (2005) offer a comprehensive introduction to FDA. A key issue in modeling functional data is the representation of the underlying process X, which is often of a complex nature and requires regularization. A common approach is to utilize functional principal component (FPC) analysis (FPCA), exploiting a data-driven eigenbasis to represent X. This has been studied extensively by Rice and Silverman (1991); James et al. (2000); Yao et al. (2005); Hall and Hosseini-Nasab (2006); Hall et al. (2006), and references therein. The eigenbasis is the unique canonical basis leading to a generalized Fourier series, i.e., the Karhunen-Loève expansion. The advantage of this expansion is that it gives the most rapidly convergent representation of X in the L^2 sense (Ash and Gardner 1975).

However, the aforementioned works on FPC approaches mostly deal with independent subjects. Very little work has appeared involving the analysis of correlated subjects or of clusters. Due to the difficulty in appropriate modelling of complex dependence structures, existing work on feasible models for correlated functional data has usually been motivated in the context of specific applications. For instance, Peng and Paul (2011) adopted a separable covariance structure for weakly correlated functional data, e.g., for growth profiles from different locations in agricultural land, while Zhou et al. (2010) considered spatially correlated FPC analysis by coupling linear mixed effects (LME) models with penalized splines. In this paper, we propose a functional data model for family-wise related individuals. Our proposal models the genetic and environmental processes both at subject level, and allows for genetic dependencies introduced by varied familial associations. This is distinct from hierarchical or multilevel FPCA (Morris et al. 2003; Di et al. 2011), where the assumptions on the within-family covariance do not allow for a variety of familial relationships.

1.1 Motivating Application

Our motivating example concerns the growth (in kilograms) as a function of age (in days) of half-sibling cows in fifteen independent families. A key issue in the analysis is the incorporation of genetic information that helps researchers understand how selective breeding can change the physical traits passed down to future generations. This understanding has economic consequences, as accurate estimation of the genetic component of an individual's trait can lead to better breeding decisions. Even small improvements in breeding practices can greatly increase food production. However, the estimation of the genetic component is complicated by the fact that it is unobservable and must be inferred from the observed physical trait. The physical trait depends not only on the genotype but also on the environmental effect, which includes factors such as habitat or food availability. Fortunately, genetic theory makes inference possible when data include information from related individuals.

This data set was first analyzed using a multivariate approach in Meyer (1985) and later, with a random regression approach for individual growth in Meyer and Hill (1997). The random regression approach uses a basis expansion with an individual's coefficients modeled as random effects. Statistical analysis is implemented with an LME model, see Demidenko (2004) and references therein for a general treatment of the random regression model using LME. However, in random regression, the choice of pre-specified basis functions is not straightforward. Although splines (in particular B-splines) have been a popular option, simulation studies in Griswold et al. (2008) indicated that B-splines do not necessarily perform well in many realistic settings. This might be caused by the "one-size-fits-all" character of B-splines, which may result in needing a fairly large number of B-spline functions. A natural approach to constructing a parsimonious model is to exploit the FPCA technique to find a data-adaptive eigenbasis, which often requires only a few leading eigenfunctions to adequately reconstruct trajectories.

1.2 Overview of the Paper

The main contribution of this paper is to develop a new FPCA framework that effectively takes into account genetic information and can be used in a variety of biological applications. The key is to generalize the canonical eigenbasis model to genetically related subjects within independently sampled families. As the individual phenotype is irregularly and sparsely observed with noise, a common occurrence in many settings, it is desirable to borrow strength from the whole sample. Yao et al. (2005) proposed a version of FPC analysis, called Principal components Analysis through Conditional Expectation (PACE), that is particularly useful for such sparse functional data. Compared to spline-based FPC methods that implicitly treat truncated models as the target (James et al. 2000), PACE emphasizes genuine nonparametric modeling of the covariance and finds data-driven eigenfunctions to be used as basis functions. Thus PACE allows for theoretical investigation of the underlying process itself. Given these advantages of the PACE approach, we couple the PACE principle with the genetic information to develop a novel FPCA framework, called Familial principal components Analysis through Conditional Expectation (FACE). Our approach naturally decomposes the total covariance into genetic and environmental components, both of which are estimated by smoothing techniques. Data-adaptive eigen-components associated with both covariance structures are obtained and used in the proposed FACE estimation of the genetically related individuals.

The remainder of this article is organized as follows. In Section 2, we introduce biological modeling of the genetic component of a physical trait, and motivate the proposed FPC model for related individuals. Section 3 describes the methodology for estimation of the model components, including the genetic and environmental covariances and their respective eigen-components. The known familial genetic relationship is utilized and leads to the proposed FACE estimation for subject-level signal extraction. We analyze the growth of beef cattle in Section 4, while Section 5 contains simulation examples. Concluding remarks are offered in Section 6.

2. GENETIC RELATIONSHIP AND PROPOSED FUNCTIONAL MODEL

2.1 Background on the Quantitative Genetic Model

To describe the standard quantitative genetic model for physical traits, let X_j denote the phenotype of individual j, Y_j the phenotype observed with error ε_j, g_j the genetic component, and e_j the environmental factor. Suppose for now that these quantities are either all scalar, p-vectors, or functions. The simplest genetic model is an additive structure with g_j, e_j, and ε_j uncorrelated with expected values equal to 0,

$$Y_j = X_j + \varepsilon_j = \mu + g_j + e_j + \varepsilon_j. \tag{1}$$

Individuals raised in different environments have uncorrelated e_j's, while related individuals from the same family have correlated underlying genotypes, the g_j's, with the amount of correlation depending on the individuals' relationship. For instance, suppose that g_j is a p-vector with $p \times p$ covariance matrix G. The $p \times p$ cross-covariance matrix defined as $\mathbb{E}[g_j g_{j'}^\top]$, $j \neq j'$, is equal to $\alpha_{jj'}G$, where $\alpha_{jj'} \in [0, 1]$ is referred to as the relationship coefficient that depends on the relationship between individuals j and j' and is twice of the entries in a so-called kinship matrix (Lynch and Walsh 1998). If the individuals are full siblings, i.e., they have the same mother and father, then $\alpha_{jj'} = 1/2$. If the individuals are half-siblings, that is, if they have only one parent in common, then $\alpha_{jj'} = 1/4$. If the individuals are unrelated then $\alpha_{jj'} = 0$, and if they are clones or the same individual then $\alpha_{jj'} = 1$. The intuition behind the value of $\alpha_{jj'}$ is that $\alpha_{jj'}$ equals the expected proportion of genes that individuals j and j' share via inheritance.

This model for genetic correlation and the use of these values of $\alpha_{jj'}$ are well-supported by both theoretical calculations and empirical studies. Their use is standard in animal breeding and in laboratory experiments in evolutionary biology. The model was first introduced, with values of $\alpha_{jj'}$ calculated, in Fisher (1918). Also see Lynch and Walsh (1998, Chapter 7) for a modern treatment

and Heckman (2003) for a statistician-friendly derivation of $\mathbb{E}[g_j g_{j'}^\top] = G/2$ for a mother-child relationship. Analysis of (1) is straightforward when the traits are scalar or vector-valued, the relationships are all the same and the design is balanced – for instance, for data from N independent families, with k full siblings in each family. In this case, variance/covariance parameters are easily estimated in closed form by analysis of variance and method of moments. For more general designs and combinations of relationships, numerical estimation is possible via (restricted) maximum likelihood (Lynch and Walsh 1998, Chapter 27), and is implemented in software such as ASReml (http://www.vsni.co.uk/software/asreml) and WOMBAT (Meyer 2007).

2.2 Functional Data Model for Genetically Related Individuals

Data such as weights of cows can be viewed as arising from smooth functions, even if the weights are sampled at irregular and, possibly, sparse discrete times across subjects. We consider the situation where there are N independent families with n_i members in family i. Let $\alpha_{i,jj'}$ denote the known relationship coefficient for individuals j and j' of family i and assume that the within-family relationship coefficients are non-zero. While our methodology holds for general $\alpha_{i,j,j'}$'s, in the data we analyze in Section 4, all family members are half-siblings, i.e., $\alpha_{i,jj'} = 1/4$ for $j \neq j'$ and $\alpha_{i,jj} = 1$ otherwise.

The functional version of (1) for the phenotype of the jth individual in the ith family is

$$X_{ij}(t) = \mu(t) + g_{ij}(t) + e_{ij}(t), \tag{2}$$

where μ is the population mean curve, g_{ij} is what is called the random genetic effect, and e_{ij} models any other random effects (mainly environmental) giving rise to within individual covariances that are not due to g_{ij}. As is common (see, e.g., Lynch and Walsh 1998), we will refer to e_{ij} as the environmental effect and g_{ij} simply as the genetic effect. In this model, g_{ij} and e_{ij} are (i) mean zero with the variance of $g_{ij}(t)$ and $e_{ij}(t)$ finite for all t, (ii) uncorrelated, (iii) $\mathrm{cov}(g_{ij}(s), g_{ij}(t)) = G(s,t)$, and (iv) $\mathrm{cov}(e_{ij}(s), e_{ij}(t)) = E(s,t)$. These four properties imply that the total covariance is

$\text{cov}(X_{ij}(s), X_{ij}(t)) = V(s,t) = G(s,t) + E(s,t)$. The within-family genetic correlation between two individuals depends on G and the individuals' relationship coefficient:

$$\text{cov}(g_{ij}(s), g_{ij'}(t)) = \alpha_{i,jj'} G(s,t). \tag{3}$$

The processes $e_{ij}(\cdot)$ and $e_{i'j'}(\cdot)$ are independent when $(i, j) \neq (i', j')$. Assume that the measurements are taken on a closed and bounded interval τ, i.e., $t \in \tau$. Note that model (2) is not the classical functional model that assumes that data come from independent realizations of $X_{ij}(t) = \mu(t) + v_{ij}(t)$. In (2), we have decomposed the random deviation $v_{ij}(t)$ as $g_{ij}(t) + e_{ij}(t)$, where the genetic effect $g_{ij}(t)$ induces a within-family correlation.

A stochastic process with finite covariance admits a Karhunen-Loève expansion and its covariance function admits a spectral basis expansion (Loève 1978; Adler and Taylor 2007). The key proposal is to exploit such expansions for both genetic and environmental processes, whilst maintaining the dependence structure of related individuals. For the genetic process g_{ij}, we have for $s, t \in \tau$,

$$g_{ij}(t) = \sum_{l=1}^{\infty} \xi_{ijl}\phi_l(t), \qquad G(s,t) = \sum_{l=1}^{\infty} \lambda_l \phi_l(s)\phi_l(t), \tag{4}$$

where the ϕ_l's are orthonormal eigenfunctions, $\xi_{ij1}, \xi_{ij2}, \ldots$ are the FPC scores, which are uncorrelated random variables with zero mean and variances $\lambda_1 > \lambda_2 > \ldots$, satisfying $\sum_{l=1}^{\infty} \lambda_l < \infty$. Based on the underlying genetic model in equation (3), we can deduce that the correlation between ξ_{ijl} and $\xi_{i'j'l'}$ is $\lambda_l\ \alpha_{i,jj'}$ for $i = i'$ and $l = l'$, and zero otherwise. This genetic association is the key to consistent parameter estimation, as it enables us to borrow information across related individuals. This model and basis expansion in the context of selection and genetics was first described in Kirkpatrick and Heckman (1989). Similar expansions hold for the environmental process e_{ij} with orthonormal eigenfunctions $\{\psi_m\}$ and nonincreasing eigenvalues $\{\rho_m\}$, i.e., for $s, t \in \tau$,

$$e_{ij}(t) = \sum_{m=1}^{\infty} \zeta_{ijm}\psi_m(t), \qquad E(s,t) = \sum_{m=1}^{\infty} \rho_m \psi_m(s)\psi_m(t), \tag{5}$$

where ζ_{ijm} are uncorrelated FPC scores of e_{ij} with zero mean and finite variance ρ_m. It is obvious that the correlation between ζ_{ijm} and $\zeta_{i'j'm'}$ is always zero given independent environmental processes, unless $(i, j, m) = (i', j', m')$.

Therefore the proposed FPC model for $X_{ij}(t)$ based on these Karhunen-Loève expansions is given by

$$X_{ij}(t) = \mu(t) + \sum_{l=1}^{\infty} \xi_{ijl}\phi_l(t) + \sum_{m=1}^{\infty} \zeta_{ijm}\psi_m(t), \ t \in \tau. \tag{6}$$

The deviation of each curve X_{ij} from the overall trend μ is a sum of curves ϕ_l and ψ_m with random amplitudes ξ_{ijl} and ζ_{ijm}, respectively. Although the underlying model (6) is infinite-dimensional, the typically rapid decay of eigenvalues often allows us to use a small number of leading eigenfunctions to recover X_{ij}. In practice, the infinite sums in (6) can be truncated and the ϕ_l's and ψ_m's estimated, yielding a data-adaptive low-dimensional model for X_{ij}. The practical choice of the level of truncations is discussed in Section 3. This eigenfunction approach differs from a random regression model with spline basis functions, as the eigenfunction basis is completely data-driven, while the spline function basis is pre-specified without knowledge of the data. A principal components approach to model (2) appears in Di et al. (2011), but with a more restricted covariance structure, which in our context would require that $\alpha_{i,jj'} \equiv \alpha$ for all i and for all $j \neq j'$.

We let the data observed for individual j from family i consist of n_{ij} repeated measurements of X_{ij} taken at discrete time points $\{T_{ijk} \in \tau : k = 1 \ldots, n_{ij}\}$. Denoting the kth observation of X_{ij} at T_{ijk} by Y_{ijk}, the data model is

$$\begin{aligned} Y_{ijk} &= X_{ij}(T_{ijk}) + \varepsilon_{ijk} \\ &= \mu(T_{ijk}) + \sum_{l} \xi_{ijl}\phi_l(T_{ijk}) + \sum_{m} \zeta_{ijm}\psi_m(T_{ijk}) + \varepsilon_{ijk}, \end{aligned} \tag{7}$$

where the ε_{ijk}'s are independent and identically distributed errors with zero mean, finite variance σ^2, and are independent of both the ξ_{ijl} and the ζ_{ijm}.

3. MODEL ESTIMATION AND FPC REPRESENTATION

The quantities in model (7) are composed of two types: the population components, such as the mean, covariances and eigenvalues/functions; and the subject-level signals, i.e., the random amplitudes or FPC scores for the underlying genetic and environmental processes. The main challenge in estimating these quantities is due to the irregularly and sparsely observed functional data. More specifically, there may be only a few observations available for some or even all of the individuals. In this case, borrowing strength across the entire collection of data is important for obtaining consistent estimation of the population quantities. As mentioned in the introduction, Yao et al. (2005) provided a thorough treatment for such sparse functional data in the case of the classical functional model with independent realizations, and proposed, namely, the PACE method. We shall generalize the key idea of PACE and take advantage of the genetic relationship (3) in model (7).

3.1 Estimation of Model Components

The mean and covariance functions are assumed to be smooth, so we can estimate them by nonparametric regression methods, which borrow information from neighboring data values. We use local linear smoothers (Fan and Gijbels 1996) for function and surface estimation. The key to estimating parameters from sparse functional data is to pool together information from all individuals, requiring the "pooled" data to be sufficiently dense. For these local smoothing steps, for a given level of smoothing we adopt the strategy of ignoring the dependency among the data from the same individual/family. However we do not ignore correlation when choosing the amount of smoothing. See Lin and Carroll (2000) for a discussion of smoothing correlated data. Automatic bandwidth choices for the amount of smoothing of functional data are available [see Rice and Silverman (1991) for leave-one-curve-out cross-validation and Müller and Prewitt (1993) for surface smoothing], even though subjective choices are often adequate in practice. Following

the spirit of Yao et al. (2005), the mean function μ evaluated at t is estimated by minimizing $\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}}\{Y_{ijk}-\alpha_0-\alpha_1(T_{ijk}-t)\}^2K_b(T_{ijk}-t)$ with respect to $(\alpha_0,\alpha_1)^\top$ and setting $\hat{\mu}(t)$ equal to the resulting α_0. Here $K_b(\cdot)=(1/b)K(\cdot/b)$, where the kernel function K is a positive density symmetric about 0, and b is the bandwidth. Due to the genetic correlation within family, we choose b by minimizing the "leave-one-family-out" cross-validation (CV),

$$CV(b)=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}}\{Y_{ijk}-\hat{\mu}^{-(i)}(T_{ijk};b)\}^2, \tag{8}$$

where $\hat{\mu}^{-(i)}(\cdot;b)$ is the estimate of μ gotten by removing all of the ith family's data.

The estimation of the covariance functions combines smoothing and the method of moments and relies upon the following key facts. Recalling that the total covariance $V(s,t)=G(s,t)+E(s,t)$, we have

$$\begin{aligned}&\operatorname{cov}\left(Y_{ijk},Y_{ijk'}\middle|T_{ijk},T_{ijk'}\right)=V(T_{ijk},T_{ijk'})+\delta_{kk'}\sigma^2\\&\alpha_{i,jj'}^{-1}\operatorname{cov}\left(Y_{ijk},Y_{ij'k'}\middle|T_{ijk},T_{ij'k'}\right)=G(T_{ijk},T_{ij'k'}),\qquad j\neq j',\end{aligned} \tag{9}$$

where $\delta_{kk'}=1$ for $k=k'$ and 0 otherwise. We define the centered observation $Y^c_{ijk}=Y_{ijk}-\hat{\mu}(T_{ijk})$, and the raw covariance observations $C_{ijkk'}=Y^c_{ijk}Y^c_{ijk'}$. Then a two-dimensional local linear smoother is employed to estimate the overall covariance function V, with $\hat{V}$ attained by smoothing the set of all raw observations $\{C_{ijkk'}:1\le k\neq k'\le n_{ij},\,j=1,\ldots,n_i,\,j=1,\ldots,n\}$. Note that in this step we have omitted the values C_{ijkk} since we expect that these are inflated by the noise variance σ^2. This fact provides motivation for our estimate of σ^2: $\hat{\sigma}^2=|\tau_1|^{-1}\int_{\tau_1}\{\tilde{V}(t)-\hat{V}(t,t)\}dt$, where $\tilde{V}$ is obtained by smoothing (T_{ijk},C_{ijkk}) over all individuals. The region of integration, τ_1, of length $|\tau_1|$, is taken as the middle half of the whole interval τ to reduce boundary effects introduced by smoothing. To better estimate $V(s,t)$ along the "height ridge" when $s\approx t$, we adjust the estimate $\tilde{V}$ using a local quadratic smoother, see Yao et al. (2003) for details. The bandwidths that control the smoothness of $\hat{V}$ and $\tilde{V}$, respectively, are also chosen by the leave-one-family-out CV in the spirit of (8).

To estimate the genetic covariance function G, the key relationship in (9) suggests borrowing data across the entire family by constructing raw cross-covariances obtained from individuals of the same family. Define such raw cross-covariance observations adjusted for relationship coefficients $\alpha_{i,jj'}$ by $G_{ijj'kk'} = \alpha_{i,jj'}^{-1} Y_{ijk}^c Y_{ij'k'}^c$. Therefore we estimate G using a two-dimensional local linear smoother of the pooled input $\{(T_{ijk}, T_{ij'k'}, G_{ijj'kk'}) : k, k' = 1, \ldots, n_{ij}, 1 \le j \ne j' \le n_i, i = 1, \ldots, n\}$, yielding the estimate $\hat{G}$. As a consequence, the environmental covariance E is easily obtained by $\hat{E} = \hat{V} - \hat{G}$.

We suggest an optional step for updating the estimates of G and E. Note that the genetic covariance G appears in the within-individual covariance and also appears in the covariance between related individuals, coupled with the relationship coefficient, as given in (3). In our initial estimate of G, we have only used the latter type of information, the information among related individuals, that is, we have only smoothed the adjusted cross-covariances $G_{ijj'kk'} = \alpha_{i,jj'}^{-1} Y_{ijk}^c Y_{ij'k'}^c$, $j \ne j'$. In our update, we add the information on G contained *within* an individual. Specifically we use our initial estimate of E and note that for $k \ne k'$, $\mathbb{E}[C_{ijkk'} - \hat{E}(T_{ijk}, T_{ijk'})] \approx G(T_{ijk}, T_{ijk'})$. Thus we can construct $\hat{G}^*$, a new estimate of G, by smoothing the combined "data": $\{C_{ijkk'} - \hat{E}(T_{ijk}, T_{ijk'}), k \ne k'\}$ and $\{G_{ijj'kk'}, j \ne j'\}$. The estimate of the environmental covariance is also updated by $\hat{E}^* = \hat{V} - \hat{G}^*$ accordingly. In practice, when the number of observations per individual is small and/or when we have a large number of individuals per family, this updating step can often be omitted as the changes in estimates are negligible.

Estimates of the eigenfunctions and eigenvalues of G and E are obtained as solutions to the eigen-equations

$$\begin{aligned} \int_\tau \hat{G}^*(s,t)\hat{\phi}_l(s)\, ds &= \hat{\lambda}_l \hat{\phi}_l(t), \\ \int_\tau \hat{E}^*(s,t)\hat{\psi}_m(s)\, ds &= \hat{\rho}_m \hat{\psi}_m(t), \end{aligned} \tag{10}$$

subject to the orthonormal constraints $\int_\tau \hat{\phi}_l(t)\hat{\phi}_{l'}(t)dt = \delta_{ll'}$ and $\int_\tau \hat{\psi}_m(t)\hat{\psi}_{m'}(t)dt = \delta_{mm'}$. This can be implemented by discretizing the smooth covariances $\hat{G}^*$ and $\hat{E}^*$ and carrying out matrix eigen-

decomposition, as described in Rice and Silverman (1991). However, the smoothed covariance functions $\hat{G}^*$ and $\hat{E}^*$ are not necessarily non-negative definite. A simple modification is to set negative estimated eigenvalues to zero, and reconstruct G and E based on (4) and (5), i.e.,

$$\begin{aligned}\tilde{G}(s,t) &= \sum_{l:\hat{\lambda}_l>0} \hat{\lambda}_l \hat{\phi}_l(s)\hat{\phi}_l(t),\\ \tilde{E}(s,t) &= \sum_{m:\hat{\rho}_m>0} \hat{\rho}_m \hat{\psi}_m(s)\hat{\psi}_m(t),\end{aligned} \tag{11}$$

which has been shown to improve the covariance estimation in terms of mean squared error (Hall et al. 2008, Theorem 1).

3.2 FPC Representation for Genetically Related Individuals

We proceed to reconstruct the individual trajectory X_{ij} in (6), which requires the estimation of the genetic and environmental FPC scores given by $\xi_{ijl} = \int_\tau \left\{X_{ij}(t) - \mu(t)\right\}\phi_l(t)dt$ and $\zeta_{ijm} = \int_\tau \left\{X_{ij}(t) - \mu(t)\right\}\psi_m(t)dt$, respectively. It is well-known that the classical integral approximation fails for sparsely observed functional data. The PACE method by Yao et al. (2005) overcomes this problem by employing the idea of the best linear unbiased prediction (BLUP) in the context of FPCA. Here we generalize the PACE method for estimating the FPC scores ξ_{ijl} and ζ_{ijm} to the case where individuals are genetically related within family. We call this generalization Familial principal component Analysis through Conditional Expectation (FACE).

In the sequel, all expectations are understood to be taken conditional on the times T_{ijk}. To calculate $\tilde{\xi}_{ijl}$, the BLUP of ξ_{ijl}, let $\boldsymbol{Y}_{ij} = (Y_{ij1},\ldots,Y_{ijn_{ij}})^\top$, $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}^\top,\ldots,\boldsymbol{Y}_{in_i}^\top)^\top$ and $N_i = \sum_{j=1}^{n_i} n_{ij}$. Recall the covariance structures in (9). Due to the genetic correlation within all individuals in family i, we infer the lth FPC score ξ_{ijl} of the genetic process g_{ij} from the observed data for all subjects in the ith family. Write the $n_{ij}\times n_{ij}$ auto-covariance matrix of $\boldsymbol{Y}_{ij}$ as $\Sigma_{i,jj} = \mathrm{cov}(\boldsymbol{Y}_{ij},\boldsymbol{Y}_{ij}) = [V(T_{ijk},T_{ijk'}) + \delta_{kk'}\sigma^2]_{1\le k,k'\le n_{ij}}$, and the $n_{ij}\times n_{ij'}$ cross-covariance matrix between $\boldsymbol{Y}_{ij}$ and $\boldsymbol{Y}_{ij'}$ by $\Sigma_{i,jj'} = \mathrm{cov}(\boldsymbol{Y}_{ij},\boldsymbol{Y}_{ij'}) = [\alpha_{i,jj'}G(T_{ijk},T_{ij'k'})]_{1\le k\le n_{ij},1\le k'\le n_{ij'}}$, where $1\le j\ne j'\le n_i$. Then we have the $N_i\times N_i$ covariance matrix of $\boldsymbol{Y}_i$, $\Sigma_{\boldsymbol{Y}_i} = \mathrm{cov}(\boldsymbol{Y}_i,\boldsymbol{Y}_i) = (\Sigma_{i,jj'})_{1\le j,j\le n_i}$. Let $\boldsymbol{\phi}_{ijl} =$

$(\phi_l(T_{ij1}), \ldots, \phi_l(T_{ijn_{ij}}))^\top$, and noting that $\alpha_{i,jj} = 1$ one has $\mathrm{cov}(\xi_{ijl}, \boldsymbol{Y}_i) = \lambda_l(\alpha_{i,j1}\boldsymbol{\phi}_{i1l}^\top, \ldots, \alpha_{i,jn_i}\boldsymbol{\phi}_{in_il}^\top)$. Finally, denote $\boldsymbol{\mu}_{ij} = (\mu(T_{ij1}), \ldots, \mu(T_{ijn_{ij}}))^\top$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^\top, \ldots, \boldsymbol{\mu}_{in_i}^\top)^\top$. By the BLUP principle, we obtain the FACE formulae for ξ_{ijl},

$$\begin{aligned}\tilde{\xi}_{ijl} &= \mathrm{cov}(\xi_{ijl}, \boldsymbol{Y}_i)\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_i)^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \\ &= \lambda_l(\alpha_{i,j1}\boldsymbol{\phi}_{i1l}^\top, \ldots, \alpha_{i,jn_i}\boldsymbol{\phi}_{in_il}^\top)\{(\Sigma_{i,jj'})_{1\le j,j\le n_i}\}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i), \end{aligned} \tag{12}$$

which is equal to $\mathbb{E}[\xi_{ill}|\boldsymbol{Y}_i]$ when all quantities are Gaussian. Substituting the estimates of model components, using the generic notation "ˆ", the FACE estimates are

$$\hat{\xi}_{ijl} = \hat{\lambda}_l(\alpha_{i,j1}\hat{\boldsymbol{\phi}}_{i1l}^\top, \ldots, \alpha_{i,jn_i}\hat{\boldsymbol{\phi}}_{in_il}^\top)\{(\widehat{\Sigma}_{i,jj'})_{1\le j,j\le n_i}\}^{-1}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i). \tag{13}$$

Since the environmental processes, the e_{ij}'s, are independent across individuals, the estimation for the FPC scores ζ_{iim} is as in PACE, i.e., only use the observed data for that subject. Denoting $\boldsymbol{\psi}_{ijm} = (\psi_m(T_{ij1}), \ldots, \psi_m(T_{ijn_{ij}}))^\top$, simple calculation by the BLUP principle yields the FACE formulae $\tilde{\zeta}_{iim}$ and its plug-in estimate $\hat{\zeta}_{ijm}$,

$$\begin{aligned}\tilde{\zeta}_{ijm} &= \rho_m\boldsymbol{\psi}_{ijm}^\top\Sigma_{i,jj}^{-1}(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij}), \\ \hat{\zeta}_{ijm} &= \hat{\rho}_m\hat{\boldsymbol{\psi}}_{ijm}^\top\widehat{\Sigma}_{i,jj}^{-1}(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_{ij}). \end{aligned} \tag{14}$$

The reconstruction of the individual trajectories is straightforward once we obtain the estimates of the FPC scores. It is customary to assume that the X_{ij}'s are well approximated by a low-dimensional expansion. Suppose we include the K_g and K_e leading eigenfunctions of g_{ij} and e_{ij} in (6), respectively, so that

$$\widehat{X}_{ij}(t) = \hat{\mu}(t) + \sum_{l=1}^{K_g}\hat{\xi}_{ijl}\hat{\phi}_l(t) + \sum_{m=1}^{K_e}\hat{\zeta}_{ijm}\hat{\psi}_m(t). \tag{15}$$

The values of K_g and K_e can be chosen by objective criteria, such as leave-one-family-out cross-validation, or the AIC based on pseudo-likelihood under Gassian assumptions in a spirit similar to

that of Yao et al. (2005). In practice, using the proportion of functional variation explained (FVE) with a suitable threshold is often satisfactory.

4. APPLICATION TO WEIGHTS OF BEEF CATTLE

The dataset we analyze here is a subset of a larger dataset used in Meyer et al. (1993) and Meyer (1999). Our data set contains weights in kilograms of 55 beef cattle from a total of 15 independent families. The cows within a family were half-siblings, having the same sire but different mothers. Thus the genetic correlation parameter $\alpha_{i,jj'} \equiv 1/4$ is known *a priori*, based on the half-sibling relationships. The phenotypic trajectories are notably irregularly and sparsely observed. The number n_i of half-siblings per family ranges from one to eight; see Figure 1(a) for the distribution of n_i's. Weighings occurred at ages ranging from 548 to 2553 days, i.e., $\tau = [548, 2553]$. The number n_{ij} of weighings per individual varied from 1 to 62, and a histogram of the n_{ij}'s is shown in Figure 1(b). Data were affected by some additional environmental factors, but for simplicity, we have not included them in our model. Including such fixed effects is, in general, straightforward, and would allow the user to model variability that is not completely due to individual effects.

The estimated mean function is shown in Figure 2, and shows, approximately, a yearly cyclical pattern that depicts the seasonal weight changes of beef cattle. The non-negative definite covariance estimates (11) for the genetic and environmental processes are shown in Figure 3(a) and 3(b), with caution when interpreting some large values in boundaries. We see that the genetic covariance is not as strong as the environmental covariance. Indeed, the environmental process explains about five and a half times the variability as the genetic process, where the surface of $G/(G+E)$ is presented in Figure 3(c) for visualizing the genetic contribution. However, the two covariances do exhibit similar patterns, with relatively high variation at late times. Another observation is that the environmental covariance seems to increase over time, which is not surprising as environmental influences may accumulate as the cows age. We chose $K_g = 3$ genetic principal components and

$K_e = 4$ environmental principal components as they explained 98% (62.5%, 29.9% and 5.6%, respectively) and 98.3% (81.6%, 8.1%, 5.2% and 3.4%, respectively) of the genetic and environmental variation. The estimated genetic and environmental eigenfunctions are given in Figures 4(a) and 4(b), respectively. From the first two eigenfunctions in each panel, one can see that the dominant variation in the genetic process concentrates around 2000 days and includes a contrast between weights at 1200 days and at 2300 days. The environmental effect shows a more constant influence over time with an early slow increase followed by a sharp drop after 2000 days (or vice versa). The updating step of the genetic and environmental covariances did not alter the estimates obviously and was not needed for this analysis.

We are primarily interested in predicting the growth of beef cattle from sparsely observed measurements. It is thus informative to assess the proposed method by comparing it with the PACE method that treats all individuals independently, i.e., that doesn't take familial genetic correlation into account. We calculate the leave-one-family-out cross-validation error given by $\sum_{i,j,k}\{Y_{ijk} - \widehat{X}_{ij}^{-i}(T_{ijk})\}^2$, where $\widehat{X}_{ij}^{-i}$ is the predicted phenotype of the jth cow in the ith family. Specifically, the model components are estimated based on data excluding family i using the method described in Section 3.1. Then the FPC scores $\hat{\xi}_{ijl}^{-i}$ and $\hat{\zeta}_{ijm}^{-i}$ are obtained by substituting these leave-one-family-out estimates, $\hat{\mu}^{-i}, \hat{\lambda}_l^{-i}, \hat{\rho}_m^{-i}, \hat{\phi}_l^{-i}, \hat{\psi}_m^{-i}, \Sigma_{i,jj'}^{-i}$, into (13) and (14), leading to $\widehat{X}_{ij}^{-i}$. We use K_g^{-i} and K_e^{-i} leading eigenfunctions, chosen to explain 98% of, respectively, the genetic and the environmental functional variation in the data. The reconstruction using the PACE method is obtained in a similar manner. See Yao et al. (2005) for details. Not surprisingly, the proposed FACE method considerably improves upon the PACE method by around 18%. Shown in Figure 5 are the cross-validated trajectory estimates for offsprings of two of the fifteen families using FACE and PACE methods. We observe that FACE offers improved predictions for these eight cows.

5. SIMULATED EXAMPLES

To further illustrate the performance of the proposed method, we carry out two simulation studies. For Simulation I, we closely mimic the cow data, using the same design, e.g., the same family sizes and times of weighings. The underlying model is (7) with K_g terms for the genetic component and K_e terms for the environmental component. The environmental covariance is derived from the first four estimated eigenfunctions, i.e., $K_e = 4$. In view of the importance of the genetic component, we examine three values of K_g: $K_g = 1, 2, 3$, and we use the corresponding genetic eigenfunctions estimated from the data. We use the half-sibling relationship coefficient $\alpha_{i,jj'} = 1/4$ for all i, j and $j' \neq j$. The genetic and environmental FPC scores ξ_{ijl} and ζ_{ijm} and the measurement errors ε_{ijk} are independently generated from normal distributions, respectively, using the estimated eigenvalues and error variance from the data. To focus our attention on the covariances and FPCs, we set the mean function μ to 0 in the data generation but still treat it as unknown in our analysis. For each underlying model, we generate 100 Monte Carlo samples, and produce two versions of $\widehat{X}_{ij}$, the FACE estimate that respects the familial genetic relationship, and the PACE estimate that ignores familial dependence. To select K_g and K_e, we again use a 98% threshold for the fraction of variance explained. Within each sample and for each estimation method, we calculate the integrated squared error (ISE) for the jth individual in the ith family, $\text{ISE}_{ij} = \int_\tau \{X_{ij}(t) - \widehat{X}_{ij}(t)\}^2 dt$, and the overall ISE is defined as $\text{ISE} = \sum_{i,j} \text{ISE}_{ij}$. Improvements of the proposed FACE method upon the PACE method are summarized in Table 1, which indicates a substantial improvement of 21% to 25%.

In Simulation II, we again follow model (7), but with $\mu(t) = t + \sin(2\pi t)$, $\phi_1(t) = \zeta_1(t) = -\cos(2\pi t/10)/\sqrt{5}$ and $\phi_2(t) = \zeta_2(t) = \sin(2\pi t/10)/\sqrt{5}$ and corresponding eigenvalues $\lambda_1 = 10$, $\lambda_2 = 5$ and $\rho_1 = 100$, $\rho_2 = 10$. The genetic and environmental FPC scores are generated from normal distributions, and the measurement error ε_{ijk} is from $N(0, 0.01)$. We still generate data for 15 families, but the number of siblings within family is chosen uniformly from $\{2, \ldots, 6\}$ and the number of observations per subject is chosen uniformly from $\{5, \ldots, 20\}$. The observation times

are uniformly distributed on $[0, 10]$. With 100 Monte Carlo samples, the ISE based on the FACE method incorporating genetic correlation outperformed the PACE method by 30% for the case of half-sibling families with $\alpha_{i,jj'} = 1/4$ for $j \neq j'$, and by 25% for the case of full-sibling families with $\alpha_{i,jj'} = 1/2$ for $j \neq j'$, see Table 1.

6. CONCLUSION

In this article, we propose a version of functional data analysis for trajectories of genetically related individuals from independent families. We are able to estimate various levels of variation: the genetic covariance, the environmental covariance induced by external factors, and the measurement error variance. A new method, named FACE, is proposed to take into account the familial correlation for estimating the genetic random effects. By making use of the auto-covariance function of each individual, we also develop a simple step to update estimates of the genetic and environmental covariance functions. We apply our method to study the growth over time of families of half-sibling cows. We show via data analysis and simulation studies that, for predicting underlying trajectories, our proposal improves considerably upon the existing PACE method designed for a sample of independent subjects. While our method does well on its own, it can also be part of a hybrid approach. For instance, our proposal can be used for dimension reduction, specifically to determine a handful of eigenfunctions that can then be used as basis functions in further analysis. Given the applicability of FACE for known genetic relationship, it would require a different modeling strategy to diagnose or estimate such relationship if unknown to researchers. In terms of computation, FACE typically requires about 50% more computing time than PACE, and the additional computation comes from estimating the genetic covariance $G(s, t)$ with a two-dimensional scatterplot smoother.

ACKNOWLEDGEMENTS

The research of Yao and Heckman is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). The research of Meyer is supported in part by Meat and Livestock Australia, Grant no. B.BFG.0050.

REFERENCES

Adler, R. J. and Taylor, J. E. (2007), *Random Fields and Geometry*, Springer Monographs in Mathematics, Springer.

Ash, R. B. and Gardner, M. F. (1975), *Topics in stochastic processes*, New York: Academic Press [Harcourt Brace Jovanovich Publishers], probability and Mathematical Statistics, Vol. 27.

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, Wiley Series in Probability and Statistics, Wiley.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2011), "Multilevel functional principal component analysis," *Annals of Applied Statistics*, 3, 458–488.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*, London: Chapman & Hall.

Fisher, R. A. (1918), "The Correlation Between Relatives on the Supposition of Mendelian Inheritance," *Transactions of the Royal Society of Edinburgh*, 52, 399–433.

Griswold, C., Gomulkiewicz, R., and Heckman, N. (2008), "Hypothesis testing in comparative and experimental studies of function-valued traits," *Evolution*, 62, 1229–1242.

Hall, P. and Hosseini-Nasab, M. (2006), "On properties of functional principal components analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 109–126.

Hall, P., Müller, H. G., and Wang, J. L. (2006), "Properties of principal component methods for functional and longitudinal data analysis," *The Annals of Statistics*, 34, 1493–1517.

Hall, P., Müller, H. G., and Yao, F. (2008), "Modeling sparse generalized longitudinal observations with latent Gaussian processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 730–723.

Heckman, N. (2003), "Functional data analysis in evolutionary biology," in *Recent Advances and Trends in Nonparametric Statistics*, eds. Akritas, M. G. and Politis, D. N., Elsevier, pp. 49–60.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000), "Principal component models for sparse functional data," *Biometrika*, 87, 587–602.

Kirkpatrick, M. and Heckman, N. (1989), "A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters," *Journal of Mathematical Biology*, 27, 429–450.

Lin, X. and Carroll, R. J. (2000), "Nonparametric function estimation for clustered data when the predictor is measured without/with error," *Journal of the American Statistical Association*, 95, 520–534.

Loève, M. (1978), *Probability Theory II*, vol. 46 of *Graduate Texts in Mathematics*, Springer.

Lynch, M. and Walsh, B. (1998), *Genetics and analysis of quantitative traits*, Sinauer.

Meyer, K. (1985), "Genetic parameters for dairy production of Australian Black and White cows," *Livestock Production Science*, 12, 205–219.

— (1999), "Estimates of genetic and phenotypic covariance functions for postweaning growth and mature weight of beef cows," *Journal of Animal Breeding and Genetics*, 116, 181–205.

— (2007), "WOMBAT – A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML)," *Journal of Zhejiang University Science*, 8, 815–821.

Meyer, K., Carrick, M. J., and Donnelly, B. J. P. (1993), "Genetic parameters for growth traits of Australian beef cattle from a multi-breed selection experiment," *Journal of Animal Science*, 71, 2614–2622.

Meyer, K. and Hill, W. (1997), "Estimation of genetic and phenotypic covariance functions for longitudinal or repeated records by restricted maximum likelihood," *Livestock Production Science*, 47, 185–200.

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003), "Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis," *Journal of the American Statistical Association*, 98, 573–597, with comments and a rejoinder by the authors.

Müller, H.-G. and Prewitt, K. A. (1993), "Multiparameter bandwidth processes and adaptive surface smoothing," *Journal of Multivariate Analysis*, 47, 1–21.

Peng, J. and Paul, D. (2011), "Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach," *Electronic Journal of Statistics*, 5, 1960–2003.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer, 2nd ed.

Rice, J. A. and Silverman, B. W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53, 233–243.

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel,

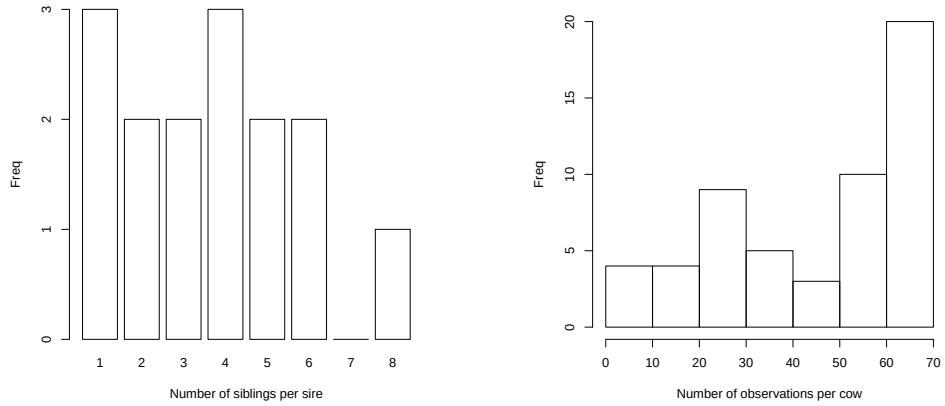
J. S. (2003), “Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate,” *Biometrics*, 59, 676–685.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), “Functional data analysis for sparse longitudinal data,” *Journal of the American Statistical Association*, 100, 577–590.

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010), “Reduced Rank Mixed Effects Models for Spatially Correlated Hierarchical Functional Data,” *Journal of the American Statistical Association*, 105, 390–400.

Table 1: ISE improvement (%), estimates of the first quartile, median, third quartile, and fraction of genetic variability to total variability of the proposed FACE method compared to PACE, where Simulation I uses data-based models with different values of (K_g, K_e) and Simulation II examines half-sibling ($\alpha = 0.25$) and full-sibling ($\alpha = 0.5$) family relationships.

	(K_g, K_e)	Mean (SE)	1st Quart.	Median	3rd Quart.	Fraction
	(1, 4)	21.4 (1.5)	15.1	23.5	28.7	30.7%
Simulation I	(2, 4)	25.1 (1.6)	12.9	28.9	36.3	33.2%
	(3, 4)	21.9 (1.6)	10.9	24.7	32.6	33.3%
	α	Mean (SE)	1st Quart.	Median	3rd Quart.	Fraction
Simulation II	0.25	30.4 (3.1)	13.4	39.0	52.8	22.6%
	0.50	25.4 (3.0)	11.7	30.4	45.4	16.7%



(a) Siblings per sire (b) Observations per cow

Figure 1: Beef cattle data: frequency distributions.

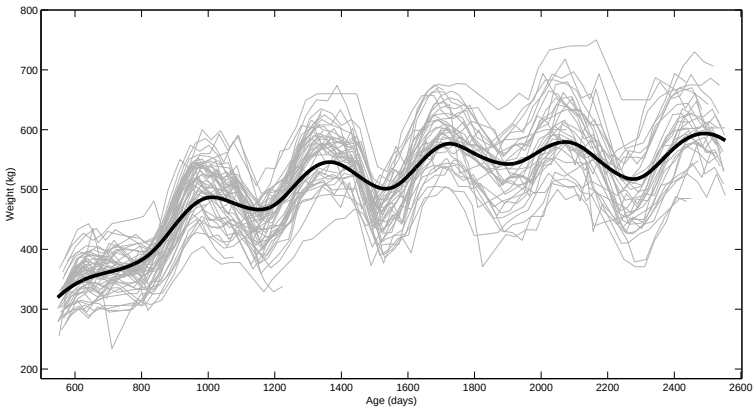
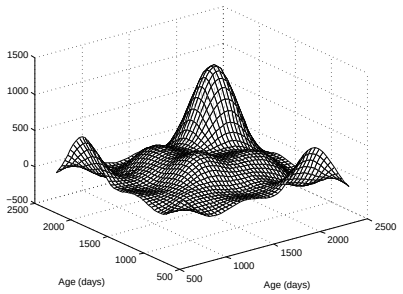
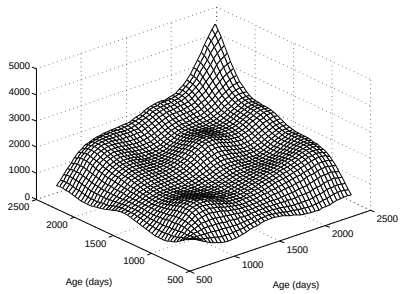


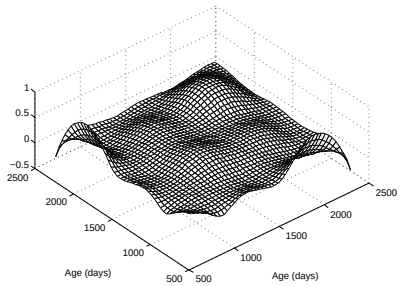
Figure 2: Estimated mean function (dark) with observed trajectories (light) for the beef cattle data.



(a) Genetic

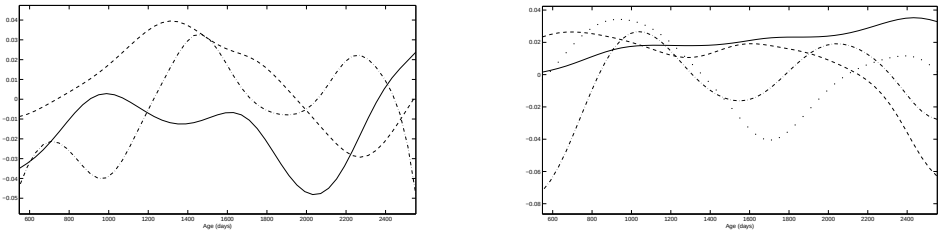


(b) Environmental



(c) Fraction

Figure 3: Non-negative definite estimates of the genetic and environmental covariance functions, as well as the fraction surface $G/(G+E)$, for the beef cattle data.



(a) Genetic (b) Environmental

Figure 4: Shown are the first (solid), second (dashed), third (dash-dot), and fourth (dotted) eigenfunctions. Left: first three eigenfunctions of the genetic process, accounting for 98% of the genetic variance. Right: first four eigenfunctions of the environmental process, explaining 98.3% of the environmental variance.

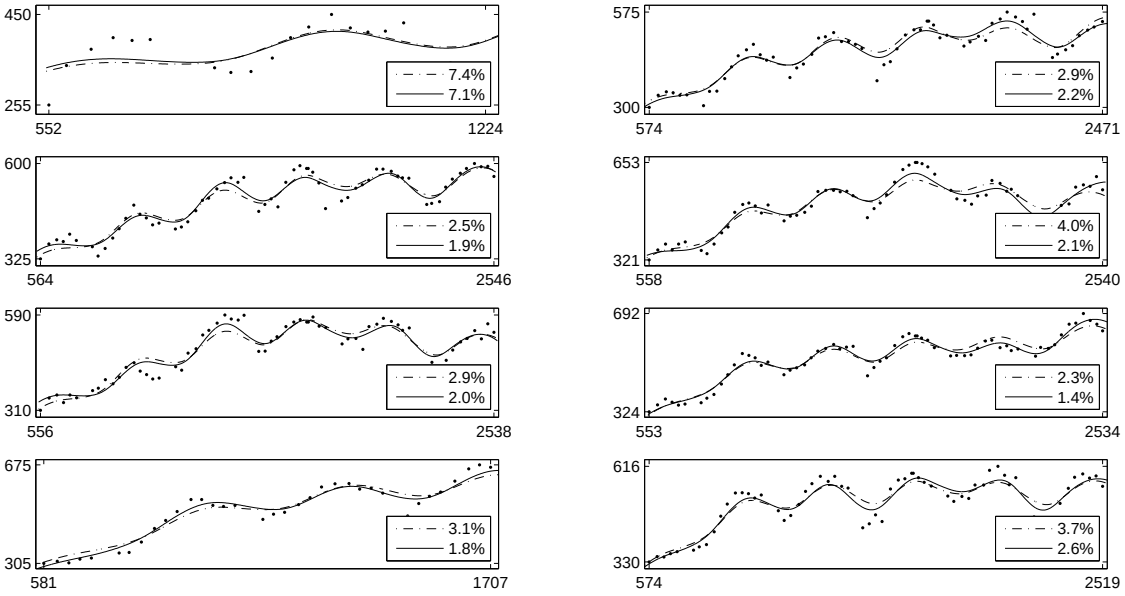


Figure 5: Estimated trajectories using leave-one-family-out cross-validation (CV) obtained using FACE method (solid) and PACE method (dashed). The data are from two families of cows; the first row presents results for two half-siblings from one family and the bottom three rows present results from six half-siblings from another family. The legend shows the relative CV error of each cow, $\sum_{k=1}^{n_{ij}}\{Y_{ijk} - \widehat{X}_{ij}^{-i}(T_{ijk})\}^2 / Y_{ijk}^2$, obtained from the two methods, where $\widehat{X}_{ij}^{-i}$ is as described in Section 4.