Penalized Clustering of Large-Scale Functional Data With Multiple Covariates

Ping MA and Wenxuan ZHONG

In this article we propose a penalized clustering method for large-scale data with multiple covariates through a functional data approach. In our proposed method, responses and covariates are linked together through nonparametric multivariate functions (fixed effects), which have great flexibility in modeling various function features, such as jump points, branching, and periodicity. Functional ANOVA is used to further decompose multivariate functions in a reproducing kernel Hilbert space and provide associated notions of main effect and interaction. Parsimonious random effects are used to capture various correlation structures. The mixed-effects models are nested under a general mixture model in which the heterogeneity of functional data is characterized. We propose a penalized Henderson's likelihood approach for model fitting and design a rejection-controlled EM algorithm for the estimation. Our method selects smoothing parameters through generalized cross-validation. Furthermore, Bayesian confidence intervals are used to measure the clustering uncertainty. Simulation studies and real-data examples are presented to investigate the empirical performance of the proposed method. Open-source code is available in the R package MFDA.

KEY WORDS: Clustering; EM algorithm; Functional data analysis; Mixed-effects model; Smoothing spline.

1. INTRODUCTION

With the rapid advancement in high-throughput technology, extensive repeated measurements have been taken to monitor the systemwide dynamics in many scientific investigations. A typical example is temporal gene expression studies, in which a series of microarray experiments are conducted sequentially during a biological process, for example, cell cycle microarray experiments (Spellman et al. 1998). At each time point, mRNA expression levels of thousands of genes are measured simultaneously. Collected over time, a gene's "temporal expression profile" gives the scientist some clues as to what role this gene may play during the process. A group of genes with similar profiles are often "co-regulated" or participants of a common and important biological function. Thus clustering genes into homogeneous groups is a crucial first step in deciphering the underlying mechanism. The need to account for intrinsic temporal dependency of repeated observations within the same individual renders traditional methods, such as K-means and hierarchical clustering, inadequate. By casting repeated observations as multivariate data with certain correlation structure, one ignores the time interval and time order of sampling. In addition, missing observations in the measurements yield an unbalanced design, necessitating imputation before applying multivariate approaches, such as the multivariate Gaussian clustering method (MCLUST; Fraley and Raftery 1990).

In addition to the time factor, such repeated measurements often contain other covariates, for example, replicates at each time point, species in comparative genomics studies (McCarroll et al. 2004), and treatment groups in case-control studies (Storey, Xiao, Leek, Tompkins, and Davis 2005), as well as many factors in a factorial designed experiment. Incorporation of multiple covariates adds another layer of complexity. Clustering methods that take all of these factors into account remain lacking.

Recently, nonparametric analysis of data in the form of curves (i.e., functional data) has been subject to active research. (See Ramsay and Silverman 2002, 2005 for a comprehensive treatment of functional data analysis.) A curve-based clustering method (FCM) was introduced by James and Sugar (2003) to cluster sparsely sampled functional data. Similar approaches were developed by Luan and Li (2003, 2004) and Heard, Holmes, and Stephens (2006) to analyze temporal gene expression data. Although these methods model the time factor explicitly, none is designed to accommodate additional factors. Moreover, smoothing-related parameters (e.g., knots and degrees of freedom) in these methods are the same across all clusters and must be specified a priori. Consequently, they cannot model drastically different patterns among different clusters, leading to high false-classification rates. Finally, the computational costs of these methods are very high for large-scale data.

Motivated by analysis of temporal gene expression data, we propose a flexible functional data clustering method that overcomes the aforementioned obstacles. In our proposed method, responses and covariates are linked together through nonparametric multivariate functions (fixed effects), which have great flexibility in modeling various function features, such as jump points, branching, and periodicity. Functional ANOVA is used to further decompose multivariate functions (fixed effects) in a reproducing kernel Hilbert space (RKHS) and provide associated notions of main effect and interaction (Wahba 1990; Gu 2002). Parsimonious random effects, complementing the fixed effects, are used to capture various correlation structures. The mixed-effects models are nested under a general mixture model, in which the heterogeneity of the functional data is characterized. We propose a penalized Henderson's likelihood approach for model fitting and design a rejection-controlled EM algorithm for estimation. In this EM algorithm, the E-step is followed by a rejection-controlled sampling step (Liu, Chen, and Wong 1998) to eliminate a significant number of functional observations with negligible posterior probabilities of belonging to a particular cluster from calculation in the subsequent M-step. The M-step is decomposed into the simultaneous maximization of penalized weighted least squares in each cluster.

Ping Ma is Assistant Professor (E-mail: *pingma@uiuc.edu*) and Wenxuan Zhong is Assistant Professor (E-mail: *wenxuan@uiuc.edu*), Department of Statistics, University of Illinois, Champaign, IL 61820. Ma's research was supported in part by National Science Foundation grant DMS-0723759. The authors are grateful to Jun S. Liu, Chong Gu, Yu Zhu, Xuming He, and Steve Portnoy for many illuminating discussions on this article, and Kurt Kwast for providing the yeast microarray data. The authors also thank the editor, the associate editor, the two referees, John Marden, and Adam Martinsek for their constructive comments and suggestions that have led to significant improvement in this article.

Journal of the American Statistical Association
June 2008, Vol. 103, No. 482, Theory and Methods
DOI 10.1198/016214508000000247

The smoothing parameters associated with the penalty are selected by generalized cross-validation (GCV), which can be shown to track a squared error loss asymptotically. Thus our method is data-adaptive and automatically captures some important functional fluctuations. For model selection, we use the Bayesian information criterion to select the number of clusters. Moreover, the proposed method provides not only subject-to-cluster assignment, but also the estimated mean function and associated Bayesian confidence intervals for each cluster. The Bayesian confidence intervals are used to measure the clustering uncertainty. These nice features make the proposed method very powerful for clustering large-scale functional data.

The remainder of the article is organized as follows. In Section 2 we present a nonparametric mixed-effects model representation for functional data. We consider a mixture model for clustering in Section 3, and simulation and real data analysis in Sections 4 and 5. A few remarks in Section 6 conclude the article. Proofs of the theorems are collected in Appendix.

2. NONPARAMETRIC MIXED–EFFECTS REPRESENTATION OF HOMOGENEOUS FUNCTIONAL DATA

Assuming that the data are homogeneous (i.e., the number of clusters is one), we present a mixed-effects representation of functional observations.

2.1 Model Specification

We assume that the functional data of the ith individual, $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, follows the mixed-effects model,

$$\mathbf{y}_i = \mu(\mathbf{x}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where the population mean μ is assumed to be a smooth function defined on a generic domain Γ, $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i})^T$ is an ordered set of sampling points, $\mathbf{b}_i \sim \mathrm{N}(0, \mathbf{B})$ is a $p \times 1$ random-effects vector associated with a $n_i \times p$ design matrix $\mathbf{Z}_i$, and random errors $\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma^2\mathbf{I})$ are independent of $\mathbf{b}_i$'s and of each other. The random-effects covariance matrix $\mathbf{B}$ and random error variance σ^2 are to be estimated from the data. Model (1) has been studied extensively in the statistical literature (see Wang 1998; Zhang, Lin, Raz, and Sowers 1998; Gu and Ma 2005; and references therein).

For multivariate x, where $x = (x_{\langle 1\rangle}, x_{\langle 2\rangle}, \ldots, x_{\langle d\rangle})^T$, each entry $x_{\langle k\rangle}$ takes values in some fairly general domain Γ_k, that is, $\Gamma = \bigotimes_{k=1}^{d} \Gamma_k$. Some examples are as follows.

Example 1. $\Gamma = [0, \mathcal{T}] \times \{1, \ldots, c\}$ to model temporal variation from time 0 to time $\mathcal{T}$ under multiple conditions, and $\Gamma = \text{circle} \times \{1, \ldots, s\}$ to model periodicity of a biological process of multiple species.

The functional ANOVA decomposition of a multivariate function μ is

$$\mu(x) = \mu_0 + \sum_{j=1}^{d} \mu_j\big(x_{\langle j\rangle}\big) + \sum_{j=1}^{d}\sum_{k=j+1}^{d} \mu_{jk}\big(x_{\langle j\rangle}, x_{\langle k\rangle}\big) + \cdots + \mu_{1,\ldots,d}\big(x_{\langle 1\rangle}, \ldots, x_{\langle d\rangle}\big), \tag{2}$$

where μ_0 is a constant, μ_j's are the main effects, μ_{jk}'s are the two-way interactions, and so on. The identifiability of the terms in (2) is ensured by side conditions through averaging operators (see Wahba 1990; Gu 2002).

By using different specifications of the random effect $\mathbf{b}_i$ and associated design matrix $\mathbf{Z}_i$, model (1) can accommodate various correlation structures.

Example 2. If we let $p = 1$ (i.e., $\mathbf{b}_i$ is a scalar, $\mathbf{B} = \sigma_b^2$, and $\mathbf{Z}_i = \mathbf{1}$), then we have the same correlation across time. If we let $p = 2$ [i.e., $\mathbf{b}_i = (b_{i1}, b_{i2})^T$, $\mathbf{B} = \begin{pmatrix} \sigma_{b_1}^2 & \sigma_{b_1b_2}^2 \\ \sigma_{b_1b_2}^2 & \sigma_{b_2}^2 \end{pmatrix}$, and $\mathbf{Z}_i = (\mathbf{1}, \mathbf{x}_i)$], then the difference between the ith subject profile and the mean profile is a linear function in time. The covariance between expression values at x_1 and x_2 for the same individual is $\sigma_{b_1}^2 + (x_1 + x_2)\sigma_{b_1b_2}^2 + x_1x_2\sigma_{b_2}^2$.

2.2 Estimation

Model (1) is estimated using penalized least squares through the minimization of

$$\sum_{i=1}^{n}(\mathbf{y}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i)^T(\mathbf{y}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i) + \sum_{i=1}^{n}\sigma^2\mathbf{b}_i^T\mathbf{B}^{-1}\mathbf{b}_i + N\lambda M(\mu), \tag{3}$$

for $N = \sum_i n_i$, where the first term measures the fidelity of the model to the data, $M(\mu) = M(\mu, \mu)$ is a quadratic functional that quantifies the roughness of μ, and λ is the smoothing parameter that controls the trade-off between the goodness of fit and the smoothness of μ. Expression (3) also is referred to as the penalized Henderson's likelihood because the first two terms are proportional to the joint density (i.e., Henderson's likelihood) of $(\mathbf{y}_i, \mathbf{b}_i)$ (Robinson 1991).

To minimize (3), we need only consider smooth functions in the space $\{\mu : M(\mu) < \infty\}$ or a subspace therein. As a abstract generalization of the vector spaces used extensively in multivariate analysis, Hilbert spaces inherit many nice properties of the vector spaces; however, they are too loose to use for functional data analysis, because even the evaluation functional $[x](f) = f(x)$, the simplest functional that may be encountered, is not guaranteed to be continuous in a general Hilbert space. For example, in the Hilbert space of square-integrable functions defined on $[0, 1]$, evaluation is not even well defined. Consequently, we can focus on a constrained Hilbert space for which the evaluation functional is continuous. Such a Hilbert space is referred to as a *reproducing kernel Hilbert space* (RKHS), for which Ramsay and Silverman (2005) suggested a nickname, continuous Hilbert space. For example, the space of functions with square-integrable second derivatives is an RKHS if it is equipped with appropriate inner products (Gu 2002). For the evaluation functional $[x](\cdot)$, by the Riesz representation theorem, there exists a nonnegative definite bivariate function $R(x, y)$, the reproducing kernel, that satisfies $\langle R(x, \cdot), f(\cdot)\rangle = f(x)$, called the "representer" of $[x](\cdot)$, in an RKHS. Given an RKHS, we may derive the reproducing kernel from the Green function associated with the quadratic functional $M(\mu)$. (The construction of the reproducing kernel is beyond the scope of this article; see Wahba 1990 and Gu 2002 for details.)

The minimization of (3) is performed in a RKHS $\mathcal{H} \subseteq \{\mu : M(\mu) < \infty\}$ in which $M(\mu)$ is a square seminorm. To incorporate (2) in estimating multivariate functions, we consider $\mu_j \in \mathcal{H}_{\langle j\rangle}$, where $\mathcal{H}_{\langle j\rangle}$ is a RKHS with tensor sum decomposition $\mathcal{H}_{\langle j\rangle} = \mathcal{H}_{0\langle j\rangle} \oplus \mathcal{H}_{1\langle j\rangle}$, where $\mathcal{H}_{0\langle j\rangle}$ is the finite-dimensional "parametric" subspace consisting of parametric functions and $\mathcal{H}_{1\langle j\rangle}$ is the "nonparametric" subspace consisting of smooth functions. The induced tensor product space is

$$\begin{aligned}\mathcal{H} &= \bigotimes_{j=1}^{d} \mathcal{H}_{\langle j\rangle} = \bigoplus_{\mathcal{S}} \left[\left(\bigotimes_{j\in\mathcal{S}} \mathcal{H}_{1\langle j\rangle} \right) \otimes \left(\bigotimes_{j\notin\mathcal{S}} \mathcal{H}_{0\langle j\rangle} \right) \right] \\ &= \bigoplus_{\mathcal{S}} \mathcal{H}_{\mathcal{S}},\end{aligned}$$

where the summation runs over all subsets $\mathcal{S} \subseteq \{1, \ldots, d\}$. The subspaces $\mathcal{H}_{\mathcal{S}}$ form two large subspaces, $\mathcal{N}_M = \{\eta : M(\mu) = 0\}$, which is the null space of $M(\mu)$, and $\mathcal{H} \ominus \mathcal{N}_M$, with the reproducing kernel $R_M(\cdot, \cdot)$. The solution of (3) has expression

$$\mu(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{T} c_i R_M(s_i, x), \tag{4}$$

where $\{\phi_\nu\}_{\nu=1}^{m}$ is a basis of $\mathcal{N}_M$, d_ν and c_i are the coefficients, and $\mathbf{s} = (s_1, \ldots, s_T)$ is a distinct combination of all x_{ij} ($i = 1, \ldots, n$, $j = 1, \ldots, n_i$).

Example 3. Consider the temporal variation under a treatments. Take the fixed effect as $\mu(t, \tau)$, where $\tau \in \{1, \ldots, a\}$ denotes the treatment levels. We can decompose

$$\mu(t, \tau) = \mu_\emptyset + \mu_1(t) + \mu_2(\tau) + \mu_{1,2}(t, \tau),$$

where $\mu_\emptyset$ is a constant, $\mu_1(t)$ is a function of t satisfying $\mu_1(0) = 0$, $\mu_2(\tau)$ is a function of τ satisfying $\sum_{\tau=1}^{a} \mu_2(\tau) = 0$, and $\mu_{1,2}(t, \tau)$ satisfies $\mu_{1,2}(0, \tau) = 0$, $\forall\tau$ and $\sum_{\tau=1}^{a} \mu_{1,2}(t, \tau) = 0$, $\forall t$. The term $\mu_\emptyset + \mu_1(t)$ is the "average variation," and the term $\mu_2(\tau) + \mu_{1,2}(t, \tau)$ is the "contrast variation."

For flexible models, we can use

$$\begin{aligned}M(\mu) = \theta_1^{-1} \int_0^{\mathcal{T}} (d^2\mu_1/dt^2)^2\, dt \\ + \theta_{1,2}^{-1} \int_0^{\mathcal{T}} \sum_{\tau=1}^{a} (d^2\mu_{1,2}/dt^2)^2\, dt,\end{aligned} \tag{5}$$

which has a null space $\mathcal{N}_M$ of dimension $2a$. A set of ϕ_ν is given by

$$\{1, t, I_{\{j\}}(\tau) - 1/a, (I_{\{j\}}(\tau) - 1/a)t,\ j = 1, \ldots, a-1\},$$

and the function R_M is given by

$$\begin{aligned}&R_M(t_1, \tau_1; t_2, \tau_2) \\ &= \theta_1 \int_0^{\mathcal{T}} (t_1 - u)_+ (t_2 - u)_+\, du \\ &\quad + \theta_{1,2}\big(I_{\{\tau_1\}}(\tau_2) - 1/a\big) \int_0^{\mathcal{T}} (t_1 - u)_+ (t_2 - u)_+\, du\end{aligned}$$

(see, e.g., Gu 2002, sec. 2.4.4). To force an additive model,

$$\mu(t, \tau) = \mu_\emptyset + \mu_1(t) + \mu_2(\tau), \tag{6}$$

which yields parallel curves at different treatments, we can set $\theta_{1,2} = 0$ and remove $(I_{\{j\}}(\tau) - 1/a)t$ from the list of ϕ_ν.

Substituting (4) into (3), we have

$$\begin{aligned}&(\mathbf{y} - \mathbf{Sd} - \mathbf{Rc} - \mathbf{Zb})^T (\mathbf{y} - \mathbf{Sd} - \mathbf{Rc} - \mathbf{Zb}) \\ &\qquad + \mathbf{b}^T \boldsymbol{\Omega} \mathbf{b} + N\lambda \mathbf{c}^T \mathbf{Q} \mathbf{c},\end{aligned} \tag{7}$$

where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$, $\mathbf{d} = (d_1, \ldots, d_m)^T$, $\mathbf{c} = (c_1, \ldots, c_T)^T$, $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)^T$, $\mathbf{S} = (\mathbf{S}_1^T, \ldots, \mathbf{S}_n^T)^T$ with the (k, ν)th entry of the $n_i \times m$ matrix $\mathbf{S}_i$ equal to $\phi_\nu(t_{ik})$, $\mathbf{R} = (\mathbf{R}_1^T, \ldots, \mathbf{R}_n^T)^T$ with the (l, j)th entry of the $n_i \times T$ matrix $\mathbf{R}_i$ equal to $R_M(t_{il}, s_j)$, the design matrix $\mathbf{Z} = \operatorname{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, $\boldsymbol{\Omega} = \sigma^2 \operatorname{diag}(\mathbf{B}^{-1}, \ldots, \mathbf{B}^{-1})$, and $\mathbf{Q}$ is a $T \times T$ matrix with the (j, k)th entry equal to $R_M(s_j, s_k)$.

Differentiating (7) with respect to $\mathbf{d}$, $\mathbf{c}$, and $\mathbf{b}$ and setting the derivatives to 0, we have

$$\begin{pmatrix} \mathbf{S}^T\mathbf{S} & \mathbf{S}^T\mathbf{R} & \mathbf{S}^T\mathbf{Z} \\ \mathbf{R}^T\mathbf{S} & \mathbf{R}^T\mathbf{R} + (N\lambda)\mathbf{Q} & \mathbf{R}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S} & \mathbf{Z}^T\mathbf{R} & \mathbf{Z}^T\mathbf{Z} + \boldsymbol{\Omega} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} \mathbf{S}^T\mathbf{y} \\ \mathbf{R}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}. \tag{8}$$

The system (8) can be solved through the pivoted Cholesky decomposition followed by backward and forward substitutions (see, e.g., Kim and Gu 2004 for details).

The fitted values $\hat{\mathbf{y}} = \mathbf{S}\hat{\mathbf{d}} + \mathbf{R}\hat{\mathbf{c}} + \mathbf{Z}\hat{\mathbf{b}}$ of (3) can be written as $\hat{\mathbf{y}} = \mathbf{A}(\lambda, \boldsymbol{\Omega})\mathbf{y}$, where $\mathbf{A}(\lambda, \boldsymbol{\Omega})$ is the smoothing matrix

$$\begin{aligned}&\mathbf{A}(\lambda, \boldsymbol{\Omega}) \\ &\quad = (\mathbf{S}, \mathbf{R}, \mathbf{Z}) \\ &\qquad \times \begin{pmatrix} \mathbf{S}^T\mathbf{S} & \mathbf{S}^T\mathbf{R} & \mathbf{S}^T\mathbf{Z} \\ \mathbf{R}^T\mathbf{S} & \mathbf{R}^T\mathbf{R} + (N\lambda)\mathbf{Q} & \mathbf{R}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S} & \mathbf{Z}^T\mathbf{R} & \mathbf{Z}^T\mathbf{Z} + \boldsymbol{\Omega} \end{pmatrix}^{+} \begin{pmatrix} \mathbf{S}^T \\ \mathbf{R}^T \\ \mathbf{Z}^T \end{pmatrix},\end{aligned} \tag{9}$$

and $\mathbf{C}^+$ denotes the Moore–Penrose inverse of $\mathbf{C}$ satisfying $\mathbf{CC}^+\mathbf{C} = \mathbf{C}$, $\mathbf{C}^+\mathbf{CC}^+ = \mathbf{C}^+$, $(\mathbf{CC}^+)^T = \mathbf{CC}^+$, and $(\mathbf{C}^+\mathbf{C})^T = \mathbf{C}^+\mathbf{C}$.

With varying smoothing parameters λ (including θ) and correlation parameters $\boldsymbol{\Omega}$, (8) defines an array of possible estimates, in which we need to choose a specific one in practice. A classic data-driven approach for selecting the smoothing parameter λ is GCV, as proposed by Craven and Wahba (1979). Treating the correlation parameters $\boldsymbol{\Omega}$ as extra smoothing parameters, we adopt the approach of Gu and Ma (2005) to estimate λ and the correlation parameters $\boldsymbol{\Omega}$ simultaneously by minimizing the GCV score,

$$V(\lambda, \boldsymbol{\Omega}) = \frac{N^{-1}\mathbf{y}^T(\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\Omega}))^2\mathbf{y}}{\{N^{-1}\operatorname{tr}(\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\Omega}))\}^2}. \tag{10}$$

Because the GCV score $V(\lambda, \boldsymbol{\Omega})$ is nonquadratic in λ and $\boldsymbol{\Omega}$, we can use standard nonlinear optimization algorithms to minimize the GCV as a function of the tuning parameters. In particular, we used the modified Newton algorithm developed by Dennis and Schnabel (1996) to find the minimizer. The distinguishing feature of GCV is that its asymptotic optimality can be justified in a decision-theoretic framework. We can define a quadratic loss function as

$$L(\lambda, \boldsymbol{\Omega}) = \frac{1}{N} \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i)^T (\hat{\mathbf{y}}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i).$$

Under general conditions, Gu and Ma (2005) showed that the GCV tracks the loss function asymptotically,

$$V(\lambda, \boldsymbol{\Omega}) - L(\lambda, \boldsymbol{\Omega}) - \frac{1}{N}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i = o_p(L(\lambda, \boldsymbol{\Omega})).$$

Note that $\boldsymbol{\epsilon}_i$ does not depend on λ and $\boldsymbol{\Omega}$. It then follows that the minimizer of the GCV score $V(\lambda, \boldsymbol{\Omega})$ approximately minimizes the loss function $L(\lambda, \boldsymbol{\Omega})$.

2.3 Bayesian Confidence Intervals

Unlike confidence estimates in parametric models, a rigorously justified interval estimate is a rarity in nonparametric functional estimation. An exception is the Bayesian confidence interval developed by Wahba (1983) from a Bayes model. A nice feature of Bayesian confidence intervals is that they have a certain across-the-function coverage property (see Nychka 1988). In this section we derive the posterior mean and variance for constructing Bayesian confidence intervals in our setting.

The regularization is equivalent to imposing a prior on the functional form of $\mu(x)$. To see this,we decompose $\mu = f_0 + f_1$, where f_0 has a diffuse prior in the space $\mathcal{N}_M$ and f_1 has an independent Gaussian process prior with mean 0 and covariance

$$E[f_1(s_k) f_1(s_l)] = \frac{\sigma^2}{N\lambda} R_M(s_k, \mathbf{s}^T)\mathbf{Q}^+ R_M(\mathbf{s}, s_l). \tag{11}$$

The minimizer of (3) can be shown to be the posterior mean under the foregoing prior by the following theorem.

Theorem 1. With the prior for μ specified earlier and a generic $np \times 1$ vector $\mathbf{z}$, the posterior mean of $\mu(x) + \mathbf{z}^T\mathbf{b}$ has expression

$$E[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}] = \boldsymbol{\phi}^T\hat{\mathbf{d}} + \boldsymbol{\xi}^T\hat{\mathbf{c}} + \mathbf{z}^T\hat{\mathbf{b}}, \tag{12}$$

where $\boldsymbol{\phi}$ is $m \times 1$ with the νth entry $\phi_\nu(\mathrm{x})$, $\boldsymbol{\xi}$ is $T \times 1$ with the ith entry $R(s_i, x)$, $\hat{\mathbf{d}}$, $\hat{\mathbf{c}}$, and $\hat{\mathbf{b}}$ are the solutions of (8).

The posterior variance is given in the following theorem.

Theorem 2. Under the model specified in Theorem 1, the posterior variance has expression:

$$\begin{aligned}
&\frac{N\lambda}{\sigma^2}\operatorname{var}[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}] \\
&\quad = \boldsymbol{\xi}^T\mathbf{Q}^+\boldsymbol{\xi} + N\lambda\mathbf{z}^T\boldsymbol{\Omega}^+\mathbf{z} + \boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\boldsymbol{\phi} \\
&\qquad - 2\boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\mathbf{R}\mathbf{Q}^+\boldsymbol{\xi} \\
&\qquad - 2N\lambda\boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\mathbf{Z}\boldsymbol{\Omega}^+\mathbf{z} \\
&\qquad - (\boldsymbol{\xi}^T\mathbf{Q}^+\mathbf{R}^T + N\lambda\mathbf{z}^T\boldsymbol{\Omega}^+\mathbf{Z}) \\
&\qquad \times \left(\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\right) \\
&\qquad \times (\mathbf{R}\mathbf{Q}^+\boldsymbol{\xi} + N\lambda\mathbf{Z}\boldsymbol{\Omega}^+\mathbf{z}),
\end{aligned}$$

where $\mathbf{W} = \mathbf{R}\mathbf{Q}^+\mathbf{R}^T + N\lambda\mathbf{Z}\boldsymbol{\Omega}^+\mathbf{Z}^T + N\lambda\mathbf{I}$.

The proofs of the foregoing two theorems are given in Appendix. Using Theorems 1 and 2, we construct the $100(1-\alpha)\%$ Bayesian confidence intervals as, $E[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}] \pm \Phi(1-\alpha/2)^{-1}\sqrt{\operatorname{var}[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}]}$, where $\Phi(1-\alpha/2)^{-1}$ is the $100(1-\alpha/2)$ percentile of the standard Gaussian distribution. Letting $\mathbf{z} = 0$, we get Bayesian confidence intervals for $\mu(x)$. Note that the construction of Bayesian confidence intervals is pointwise. Whether the across-the-function coverage property of Nychka (1988) holds in our case is unclear.

3. THE MIXTURE MODEL

Based on the mixed-effects representation of homogeneous functional data, we now present a mixture model for characterizing the heterogeneity.

3.1 The Model Specification

When the population is heterogeneous, we assume that the ith functional observation can be modeled as

$$\mathbf{y}_i = \mu_k(\mathbf{x}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad \text{with probability } p_k, \tag{13}$$

where $k = 1, \ldots, K$, the kth cluster's mean μ_k is a smooth function defined on a generic domain Γ, $\mathbf{b}_i \sim \mathrm{N}(0, \mathbf{B}_k)$ is a $p \times 1$ random-effects vector associated with a $n_i \times p$ design matrix $\mathbf{Z}_i$, $\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma^2\mathbf{I})$ are random errors independent of the $\mathbf{b}_i$'s and of one another, cluster probabilities p_k satisfy $\sum_{k=1}^{K} p_k = 1$, and K is the number of clusters in the population.

To ease the computation, we introduce a "latent" membership labeling variable J_{ik} such that $J_{ik} = 1$ indicates that individual i belongs to the kth cluster and $J_{ik} = 0$ otherwise. Thus we have the probability that $J_{ik} = 1$ is p_k. The mixture Henderson's likelihood is seen to be

$$\sum_{i=1}^{n} \log \sum_{k=1}^{K} [p_k f_y(\mathbf{y}_i; \mathbf{b}_i, J_{ik} = 1) f_b(\mathbf{b}_i; J_{ik} = 1)],$$

where f_y and f_b are probability density functions for $\mathbf{y}_i$ and $\mathbf{b}_i$.

3.2 Estimation

The negative penalized Henderson's likelihood of complete data $(\mathbf{y}_i, J_{ik})$ where $i = 1, \ldots, n$, is seen to be

$$\begin{aligned}
L_c = \text{constant} &- \sum_{i=1}^{n}\sum_{k=1}^{K} J_{ik} \log p_k \\
&+ \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K} J_{ik}\big[(\mathbf{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i)^T \\
&\times (\mathbf{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i) + \sigma^2\mathbf{b}_i^T\mathbf{B}_k^{-1}\mathbf{b}_i\big] \\
&+ \sum_{k=1}^{K} N\lambda_k M(\mu_k),
\end{aligned} \tag{14}$$

where λ_k is the smoothing parameter for μ_k.

Once the penalized Henderson's likelihood (14) is obtained, the EM algorithm (Dempster, Laird, and Rubin 1977; Green 1990) can be derived as follows: The E-step simply requires calculation of

$$w_{ik} = \frac{p_k\varphi(\mathbf{y}_i; \mu_k(\mathbf{x}_i), \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} p_l\varphi(\mathbf{y}_i; \mu_l(\mathbf{x}_i), \boldsymbol{\Sigma}_l)}, \tag{15}$$

where $\boldsymbol{\Sigma}_k = \mathbf{Z}_i\mathbf{B}_k\mathbf{Z}_i^T + \sigma^2\mathbf{I}$ and φ is the Gaussian density function.

The M-step requires the conditional minimization of the equation

$$-\sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}\log p_k \tag{16}$$

$$+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\big[(\mathbf{y}_i-\mu_k(\mathbf{x}_i)-\mathbf{Z}_i\mathbf{b}_{ik})^T \times(\mathbf{y}_i-\mu_k(\mathbf{x}_i)-\mathbf{Z}_i\mathbf{b}_{ik})+\sigma^2\mathbf{b}_{ik}^T\mathbf{B}_k^{-1}\mathbf{b}_{ik}\big]$$

$$+\sum_{k=1}^{K} N\lambda_k M(\mu_k), \tag{17}$$

where $\mathbf{b}_{ik}$ is $\mathbf{b}_i$ given the membership J_{ik}. Thus the M-step is equivalent to minimizing (16) and (17) separately.

By minimizing (16), we have

$$p_k=\frac{1}{n}\sum_{i=1}^{n} w_{ik} \quad \text{for } k=1,\ldots,K. \tag{18}$$

By minimizing (17), we can minimize the following K equations simultaneously

$$\sum_{i=1}^{n} w_{ik}\big[(\mathbf{y}_i-\mu_k(\mathbf{x}_i)-\mathbf{Z}_i\mathbf{b}_{ik})^T(\mathbf{y}_i-\mu_k(\mathbf{x}_i)-\mathbf{Z}_i\mathbf{b}_{ik}) +\sigma^2\mathbf{b}_{ik}^T\mathbf{B}_k^{-1}\mathbf{b}_{ik}\big]+N\lambda_k M(\mu_k), \qquad k=1,\ldots,K, \tag{19}$$

where $1/2\sigma^2$ is absorbed into λ_k. The minimization of (19) is performed in the RKHS $\mathcal{H}\subseteq\{\eta: M(\mu)<\infty\}$. Substituting solution (4) into (19), we have

$$(\mathbf{y}-\mathbf{S}\mathbf{d}_k-\mathbf{R}\mathbf{c}_k-\mathbf{Z}\mathbf{b}_k)^T\mathbf{W}_k(\mathbf{y}-\mathbf{S}\mathbf{d}_k-\mathbf{R}\mathbf{c}_k-\mathbf{Z}\mathbf{b}_k) +\mathbf{b}_k^T\tilde{\mathbf{W}}_k^{1/2}\boldsymbol{\Omega}_k\tilde{\mathbf{W}}_k^{1/2}\mathbf{b}_k+N\lambda_k\mathbf{c}_k^T\mathbf{Q}\mathbf{c}_k, \tag{20}$$

where $\mathbf{y}=(\mathbf{y}_1^T,\ldots,\mathbf{y}_n^T)^T$, $\mathbf{d}_k=(d_{1k},\ldots,d_{mk})^T$, $\mathbf{c}_k=(c_{1k},\ldots,c_{Tk})^T$, $\mathbf{b}_k=(\mathbf{b}_{1k}^T,\ldots,\mathbf{b}_{nk}^T)^T$, $\mathbf{S}=(\mathbf{S}_1^T,\ldots,\mathbf{S}_n^T)^T$ with the (k,v)th entry of the $n_i\times m$ matrix $\mathbf{S}_i$ equal to $\phi_v(t_{ik})$, $\mathbf{R}=(\mathbf{R}_1^T,\ldots,\mathbf{R}_n^T)^T$ with the (l,j)th entry of the $n_i\times T$ matrix $\mathbf{R}_i$ equal to $R_M(t_{il},s_j)$, the design matrix $\mathbf{Z}=\operatorname{diag}(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)$, $\mathbf{W}_k=\operatorname{diag}(w_{1k}\mathbf{I}_{n_1},\ldots,w_{nk}\mathbf{I}_{n_n})$, $\tilde{\mathbf{W}}_k=\operatorname{diag}(w_{1k}\mathbf{I}_p,\ldots,w_{nk}\mathbf{I}_p)$, $\boldsymbol{\Omega}_k=\sigma^2\operatorname{diag}(\mathbf{B}_k^{-1},\ldots,\mathbf{B}_k^{-1})$, and $\mathbf{Q}$ is a $T\times T$ matrix with the (j,k)th entry equal to $R_M(s_j,s_k)$.

Writing (20) in a more compact form, we have

$$(\mathbf{y}_{wk}-\mathbf{S}_{wk}\mathbf{d}_k-\mathbf{R}_{wk}\mathbf{c}_k-\mathbf{Z}\mathbf{b}_{wk})^T \times(\mathbf{y}_{wk}-\mathbf{S}_{wk}\mathbf{d}_k-\mathbf{R}_{wk}\mathbf{c}_k-\mathbf{Z}\mathbf{b}_{wk}) +\mathbf{b}_{wk}^T\boldsymbol{\Omega}_k\mathbf{b}_{wk}+N\lambda_k\mathbf{c}_k^T\mathbf{Q}\mathbf{c}_k, \tag{21}$$

where $\mathbf{y}_{wk}=\mathbf{W}_k^{1/2}\mathbf{y}$, $\mathbf{S}_{wk}=\mathbf{W}_k^{1/2}\mathbf{S}$, $\mathbf{R}_{wk}=\mathbf{W}_k^{1/2}\mathbf{R}$, $\mathbf{Z}_{wk}=\mathbf{W}_k^{1/2}\mathbf{Z}\tilde{\mathbf{W}}_k^{-1/2}$, and $\mathbf{b}_{wk}=\tilde{\mathbf{W}}_k^{1/2}\mathbf{b}_k$. Then (21) can be minimized using the techniques developed in Section 2.

The variance of measurement error is estimated as

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(\mathbf{y}_i-\mu_k(\mathbf{x}_i)-\mathbf{Z}_i\mathbf{b}_{ik})^T \times(\mathbf{y}_i-\mu_k(\mathbf{x}_i)-\mathbf{Z}_i\mathbf{b}_{ik}). \tag{22}$$

The algorithm iterates through (15), (18), (21), and (22) until all the parameters converge.

The selection of the smoothing parameters $\boldsymbol{\Omega}_k$ and λ_k plays an important role in the proposed algorithm. When first running our algorithm, in each iteration, we selected the optimal smoothing parameters for each cluster using GCV in (21). Once all parameters converge, we fixed the selected smoothing parameters and ran our algorithm for fixed smoothing parameters.

Once we fit the mixture model to the data, we could give a probabilistic (soft) clustering of each observation $\mathbf{y}_i$; that is, for each $\mathbf{y}_i$, $w_{i1},\ldots,w_{iK}$ give the estimated probabilities that this observation belongs to the first, second, …, and Kth components of the mixture. But in many practical settings, it is highly desirable to give hard clusterings of these observations by assigning each observation to one component of the mixture. In the rest of the article, we adopt the hard clustering of McLachlan and Peel (2001) by estimating the membership label

$$\hat{J}_{ik}=\begin{cases}1 & \text{if } k=\arg\max_h w_{ih}\\ 0 & \text{otherwise}\end{cases}$$

where $k=1,\ldots,K$ and $i=1,\ldots,n$.

3.3 Efficient Computation With Rejection Control

With thousands of observations under consideration, the E-step (15) results in a huge number of w_{ik}'s, many of which are extremely small. With the presence of these small w_{ik}'s, the calculation of matrices involved in the M-step (21) is expensive, unstable, and sometimes even infeasible. To alleviate the computation and stabilize the algorithm, we propose adding a rejection control step (Liu et al. 1998; Ma, Castillo-Davis, Zhong, and Liu 2006) in the EM algorithm and refer to the modified algorithm as the rejection-controlled EM algorithm.

First, we set up a threshold value c (e.g., $c=.05$). Given this threshold value, we introduce the following rejection controlled step:

$$w_{ik}^*=\begin{cases}w_{ik} & \text{if } w_{ik}>c\\ c & \text{with probability } w_{ik}/c \text{ if } w_{ik}\le c\\ 0 & \text{with probability } 1-w_{ik}/c \text{ if } w_{ik}\le c.\end{cases}$$

The resulting w_{ik}^* must be normalized, $w_{ik}^{**}=w_{ik}^*/\sum_k w_{ik}^*$. Then we replace w_{ik} by w_{ik}^{**} immediately after the E-step (15). Note that when $c=0$, the proposed algorithm is exactly the original EM algorithm, whereas when $c=1$, the proposed algorithm reduces to a variant of the Monte Carlo EM algorithm (Wei and Tanner 1990). In this way, it is possible to make accurate approximations during the E-step while greatly reducing the computation of the M-step.

Finally, to avoid local optima, we run the rejection-controlled EM with multiple chains. In practice, we first set the threshold c close to 1 at an early stage of the iterations to expedite the calculation, then gradually lower c so that the algorithm can achieve a better approximation of the original EM.

A critical issue arising from the new algorithm is how to choose an appropriate stopping rule. For the original EM algorithm, the likelihood function increases after each iteration, so we can stop the iteration when the likelihood does not change. But for the rejection-controlled EM algorithm, the likelihood functions fluctuates because of the sampling scheme, so a stopping rule like those used in the Gibbs sampler is used. When the likelihood function is no longer increasing for several consecutive iterations, we stop and choose the estimates with the highest likelihood.

3.4 Selection of the Number of Clusters

The success of our proposed methods depends heavily on the selection of the number of clusters K. A natural choice in model-based clustering is to use the BIC. The BIC imposes a penalty on the total number of parameters, scaled by the logarithm of sample size, so as to strike a balance between the goodness of fit and the model complexity. A critical issue when using BIC in nonparametric settings is to determine the effective number of parameters. Here we use the trace of the smoothing matrix to approximate the number of parameters in each cluster (Hastie and Tibshirani 1990; Gu 2002). Thus, under our model, the BIC is

$$BIC = -2\sum_{i=1}^{n}\log\sum_{k=1}^{K} p_k\varphi(\mathbf{y}_i;\mu_k(\mathbf{x}_i),\boldsymbol{\Sigma}_k) + \left(\sum_{k=1}^{K}\operatorname{tr}\mathbf{A}_k(\lambda_k,\boldsymbol{\Omega}_k)+P\right)\log N, \quad (23)$$

where $\mathbf{A}_k$ is the smoothing matrix for the kth cluster as defined in (9) and P is the number of free parameters in p_k, λ_k, and $\boldsymbol{\Omega}_k$, where $k = 1,\ldots,K$.

4. SIMULATION

To assess the performance of the proposed method, we carried out extensive analyses on simulated data sets.

4.1 Simulation 1

This simulation was designed to demonstrate the performance of the proposed method when the underlying clusters' mean functions are different for different clusters. First, we generated 100 replicates of samples according to

$$\begin{aligned} y_{1ij\tau} &= 3\sin(6\pi t_j)(1-t_j) + 2I_{\{1\}}(\tau) - 1 + \epsilon_{1ij\tau}, \\ &\qquad i = 1,\ldots,30; \\ y_{2ij\tau} &= 3\sin(6\pi t_j)(1-t_j) + \epsilon_{2ij\tau}, \\ &\qquad i = 1,\ldots,40; \\ y_{3ij\tau} &= 1{,}980t_j^7(1-t_j)^3 + 858t_j^2(1-t_j)^{10} - 2 + \epsilon_{3ij\tau}, \\ &\qquad i = 1,\ldots,50; \end{aligned}$$

and

$$y_{4ij\tau} = 3\sin(2\pi t_j) + 2I_{\{1\}}(\tau) - 1 + \epsilon_{4ij\tau}, \qquad i = 1,\ldots,30,$$

where $t_j = 1/15, 2/15, \ldots, 1$, $\tau = 0, 1$, indicator function $I_{\{1\}}(\tau) = 1$ if $\tau = 1$ and 0 otherwise, and random errors ϵ were generated from a Gaussian distribution with mean 0 and covariance matrix as

$$\operatorname{var}[\epsilon_{lij\tau}] = 1, \qquad \operatorname{cov}\left(\epsilon_{lij\tau_1}, \epsilon_{lik\tau_2}\right) = .2, \quad \text{for } l = 1, 3,$$

and

$$\operatorname{var}[\epsilon_{lij\tau}] = 1.2, \qquad \operatorname{cov}\left(\epsilon_{lij\tau_1}, \epsilon_{lik\tau_2}\right) = .4, \quad \text{for } l = 2, 4.$$

We analyzed the simulated data using the proposed method with the mixture model

$$\mathbf{y}_i = \mu_k(\mathbf{t},\tau) + b_i\mathbf{1} + \boldsymbol{\epsilon}_i \quad \text{with probability } p_k,$$

where $k = 1,\ldots,K$, $\tau = 1, 2$ for two groups and $b_i \sim \mathrm{N}(0,\sigma_b^2)$ is the individual specific random effect. The important feature of the simulated data is that the true mean curves in two groups, indexed by τ, are either identical or parallel. We built this information into our method by enforcing the additive model (6). We used the penalized Henderson's likelihood for estimation with roughness penalty $M(\mu) = \int_0^1 (d^2\mu_1/dt^2)^2\,dt$.

We compared our method with MCLUST (Fraley and Raftery 1990), FCM classification likelihood (FCMc), and FCM mixture likelihood (FCMm) (James and Sugar 2003). Because the number of clusters must be specified a priori in the partially implemented FCM software, we gave a significant starting advantage to the FCM algorithm by letting the number of clusters be the true number of clusters (four). For MCLUST, we report the clustering result with optimal BIC, which was estimated from eight models with different covariance structures. The estimated mean curves using the proposed method for each cluster and the true curves of one sample are plotted in Figure 1.

For comparison, we need a measure of the agreement of the clustering results with the true cluster membership. A popular one is the Rand index, the percentage of concordance pairs over all possible data pairs. Hubert and Arabie (1985) proposed an adjusted Rand index that takes 1 as the maximum value when two clustering results are the same and is expected to be 0 when two clustering results are independent. The boxplots of the adjusted Rand indices for clustering results using the different methods are presented in Figure 2(a). We found that across 100 samples, the average of the adjusted Rand indices for the proposed method is .9676 (median is .9838), whereas those of MCLUST, FCMm, and FCMc are .7553, .8936, and .8896. Moreover, the interquartile ranges of the adjusted Rand indices of the proposed method are .0565 (.0972, .2189, and .2262 for MCLUST, FCMm, and FCMc). These results suggest that the proposed method outperforms FCMc and FCMm (even under the ideal scenario where the true number of clusters is provided to FCMc and FCMm a priori), as well as MCLUST.

4.2 Simulation 2

This simulation is designed to demonstrate the performance of the proposed method when the underlying clusters' covariance structures are different for different clusters. We generated replicates of samples according to

$$\begin{aligned} y_{1ij\tau} &= 3\sin(6\pi t_j)(1-t_j) + 2I_{\{1\}}(\tau) - 1 + \epsilon_{1ij\tau}, \\ &\qquad i = 1,\ldots,30; \\ y_{2ij\tau} &= 3\sin(6\pi t_j)(1-t_j) + 2I_{\{1\}}(\tau) - 1 + \epsilon_{2ij\tau}, \\ &\qquad i = 1,\ldots,40; \\ y_{3ij\tau} &= 1{,}980t_j^7(1-t_j)^3 + 858t_j^2(1-t_j)^{10} - 2 + \epsilon_{3ij\tau}, \\ &\qquad i = 1,\ldots,50; \end{aligned}$$

and

$$y_{4ij\tau} = 1{,}980t_j^7(1-t_j)^3 + 858t_j^2(1-t_j)^{10} - 2 + \epsilon_{4ij\tau}, \qquad i = 1,\ldots,30,$$

where $t_j = 1/15, 2/15, \ldots, 1$, $\tau = 0, 1$, and indicator function $I_{\{1\}}(\tau) = 1$ if $\tau = 1$ and 0 otherwise. We generated random errors ϵ from a Gaussian distribution with mean 0 and covariance

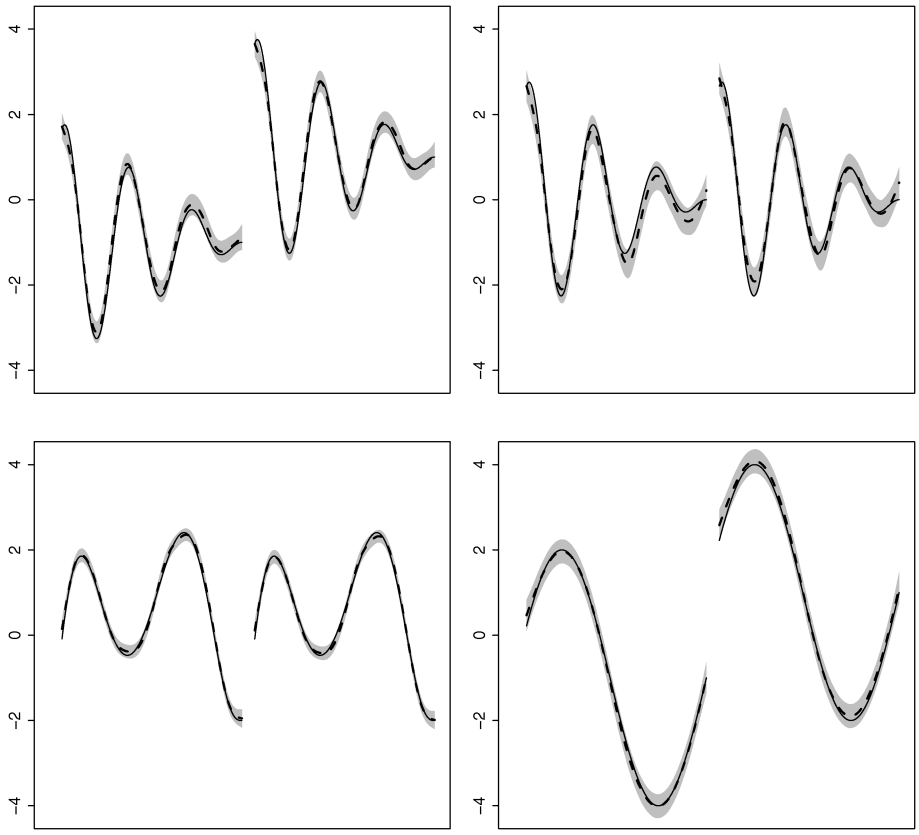

Figure 1. The estimated mean curves (- - - - -) and 95% Bayesian confidence intervals for one simulated data set. The true functions are superimposed as solid lines.

matrix

$$\text{var}[\epsilon_{lij\tau}] = 1,$$
$$\text{cov}\big(\epsilon_{lij\tau_1}, \epsilon_{lik\tau_2}\big) = \begin{cases} -.5 & \text{if } \tau_1 \neq \tau_2, \text{ for } l = 1, 3 \\ .5 & \text{otherwise} \end{cases}$$

and

$$\text{var}[\epsilon_{lij\tau}] = 1, \qquad \text{cov}\big(\epsilon_{lij\tau_1}, \epsilon_{lik\tau_2}\big) = .5 \quad \text{for } l = 2, 4.$$

We analyzed the simulated data using the proposed method with the mixture model

$$\mathbf{y}_i = \mu_k(\mathbf{t}, \tau) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad \text{with probability } p_k,$$

where $k = 1, \ldots, K$, $\tau = 1, 2$ for two groups; $\mathbf{b}_i = (b_{i1}, b_{i2})$ is the random effect with $\text{var}(b_{i1}) = \text{var}(b_{i2}) = \sigma_b^2$; and $\text{cov}(b_{i1}, b_{i2}) = \sigma_{12}^2$, $\mathbf{Z}_i = \text{diag}(\mathbf{1}_{15}, \mathbf{1}_{15})$ is a 30×2 random-effects design matrix. We enforced additive model (6). We use the pe-

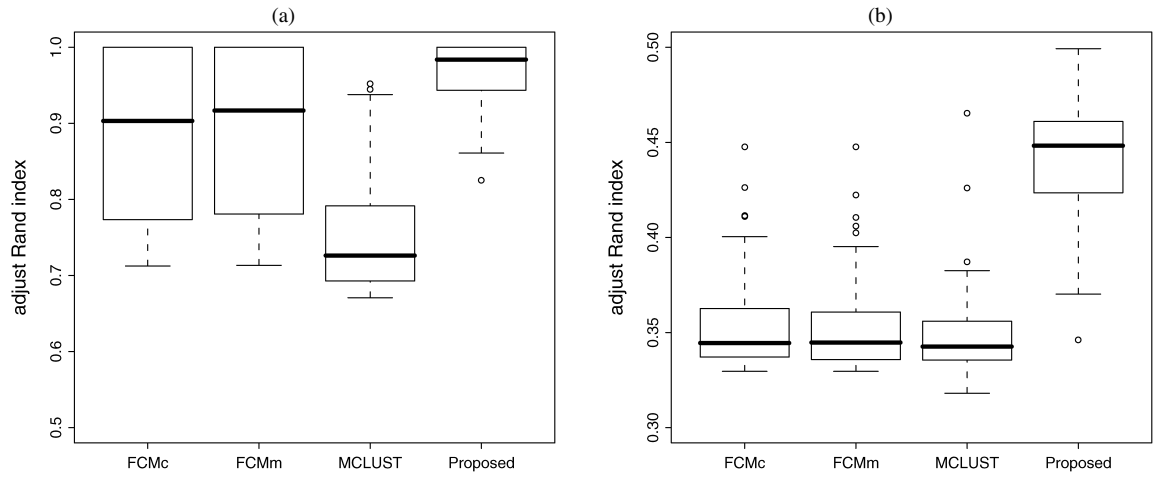

Figure 2. Boxplots of the adjusted Rand indices for clustering results of 100 simulated samples against the true clustering membership for all methods for simulations 1 (a) and 2 (b).

nalized Henderson's likelihood for estimation with roughness penalty $M(\mu) = \int_0^1 (d^2\mu_1/dt^2)^2\,dt$.

The boxplots of the adjusted Rand indices for clustering results using different methods are presented in Figure 2(b). We found that across 100 samples, the average of the adjusted Rand indices for the proposed method is .4410, whereas those of MCLUST, FCMm, FCMc are .3476, .3520, and .3530. These results suggest that the proposed method still outperforms FCMc and FCMm as well as MCLUST.

5. REAL DATA EXAMPLES

5.1 Comparative Genomic Study of Fruitfly and Worm Gene Expressions

Development is an important biological process that shares many common features among different organisms. It is well known that *D. melanogaster* (fruitfly) and *C. elegance* (worm) are two highly diverged species, the last common ancestor of which existed about 1 billion years ago. Their development is an area of active research. Arbeitman et al. (2002) measured the mRNA levels of 4,028 genes in *D. melanogaster* using cDNA microarrays during 62 time points starting at fertilization and spanning embryonic, larval, and pupal (metamorphosis) stages and the first 30 days of adulthood. mRNA was extracted from mixed male and female populations until adulthood, when males and females were sampled separately. Jiang et al. (2001) reported a cDNA microarray experiment for 17,871 genes over the life cycle of *C. elegans* at 6 time points, including eggs, larval stages (L1, L2, L3, and L4), and young adults.

To study the genomic connections in expression patterns across the two species, we combined the gene expression data sets of Arbeitman et al. (2002) and Jiang et al. (2001) using the orthologous genes provided by McCarroll et al. (2004), which resulted in a merged expression data set containing 808 orthologous genes. We analyzed the data using the proposed method with the mixture model

$$\mathbf{y}_i = \mu_k(\mathbf{t}, \tau) + b_i\mathbf{1} + \boldsymbol{\epsilon}_i,$$

with probability p_k, where $k = 1, \ldots, K$, $\tau = 1$ for fruitfly and $\tau = 2$ for worm and $b_i \sim \mathrm{N}(0, \sigma_b^2)$ is the gene-specific random effect. We used the penalized Henderson's likelihood with roughness penalty M of the form (5). We modeled sex differentiation of the fruitfly by a branching spline (Silverman and Wood 1987), the general analytic form of which with two branches on the right is

$$\mu(t) = \begin{cases} \displaystyle\sum_{v=1}^{m} d_v\phi_v(t) + \sum_{i=1}^{k} c_i R_M(s_i, t) & \text{if } t \le s_k \\ \displaystyle\sum_{v=1}^{m} d_v\phi_v(t) + \sum_{i=1}^{k} c_i R_M(s_i, t) & \\ \quad + \displaystyle\sum_{i=k+1}^{T} c_{1i} R_M(s_i - s_k, t - s_k) & \text{if } t > s_k \\ \displaystyle\sum_{v=1}^{m} d_v\phi_v(t) + \sum_{i=1}^{k} c_i R_M(s_i, t) & \\ \quad + \displaystyle\sum_{i=k+1}^{T} c_{2i} R_M(s_i - s_k, t - s_k) & \text{if } t > s_k, \end{cases} \tag{24}$$

where s_k is the branching point and the second and third rows are expressions of the two branches. A cubic smoothing spline was used. The 808 genes were clustered by our method into 34 clusters. Biological functions of genes in each cluster were annotated using Gene Ontology, and Bonferroni-corrected p values of biological function enrichment were calculated based on the hypergeometric distribution (Castillo-Davis and Hartl 2003). Of the 34 clusters discovered, 21 clusters exhibited significant biological function overrepresentation (p value $< .05$). The estimated mean gene expression curves of three clusters and their 95% Bayesian confidence intervals are given in Figure 3.

In cluster A, which consists of 31 genes, gene expressions of worms have peaks at eggs, larvae, and young adults. In the same cluster, we observed that fruitfly gene expressions that are up-regulated during embryogenesis are also up-regulated during metamorphosis, suggesting that many genes used for pattern formation during embryogenesis (the transition from egg to larva) are redeployed during metamorphosis (the transition from larva to adult). Consistently, this cluster is enriched for genes involved in embryonic development ($p = .0003$), postembryonic body morphogenesis ($p = .007$), and mRNA processing ($p = .002$), among others.

In cluster B, consisting of 24 genes, gene expressions of worms increase starting at eggs until they reach a peak at a late larval stage, then decrease during adulthood. But we observed that fruitfly gene expressions that are down-regulated during embryogenesis are up-regulated during metamorphosis and adulthood, suggesting that many genes are involved in development. The enriched gene functions are embryonic ($p = .02$), larval development ($p = .008$), and growth regulation ($p < 10^{-5}$).

Cluster C contains 25 genes. For worms, gene expressions have peaks at larva and adult stages. An overrepresentation of such gene functions as reproduction ($p < 10^{-6}$), larval development ($p < 10^{-7}$) are present in this cluster. Fruitflies show peaks in gene expression in the early embryo and in older females (but not males). An overrepresentation of such gene functions as reproduction ($p < 10^{-6}$) and embryonic development ($p < 10^{-5}$) are present in this cluster. Among related functions, this cluster also contains functions of female gamete generation, growth, and positive regulation of growth rate. Thus genes of this cluster are believed to participate in sex determination, female egg production and growth regulation.

5.2 Budding Yeast Gene Expression Under Aerobic and Anaerobic Conditions

To study the oxygen-responsive gene networks, Lai, Kosorukoff, Burke, and Kwast (2006) used cDNA microarray to monitor the gene expression changes of wild-type budding yeast (*Saccharomyces cerevisiae*) under aerobic and anaerobic conditions in a galactose medium. Under aerobic conditions, the oxygen concentration was gradually decreased until oxygen was exhausted over a 10-minute period. After 24 hours of anaerobiosis, the oxygen concentration was progressively increased back to the normal level during another of 10-minute period known as the anaerobic conditions. Microarray experiments were conducted at 14 time points under aerobic conditions and

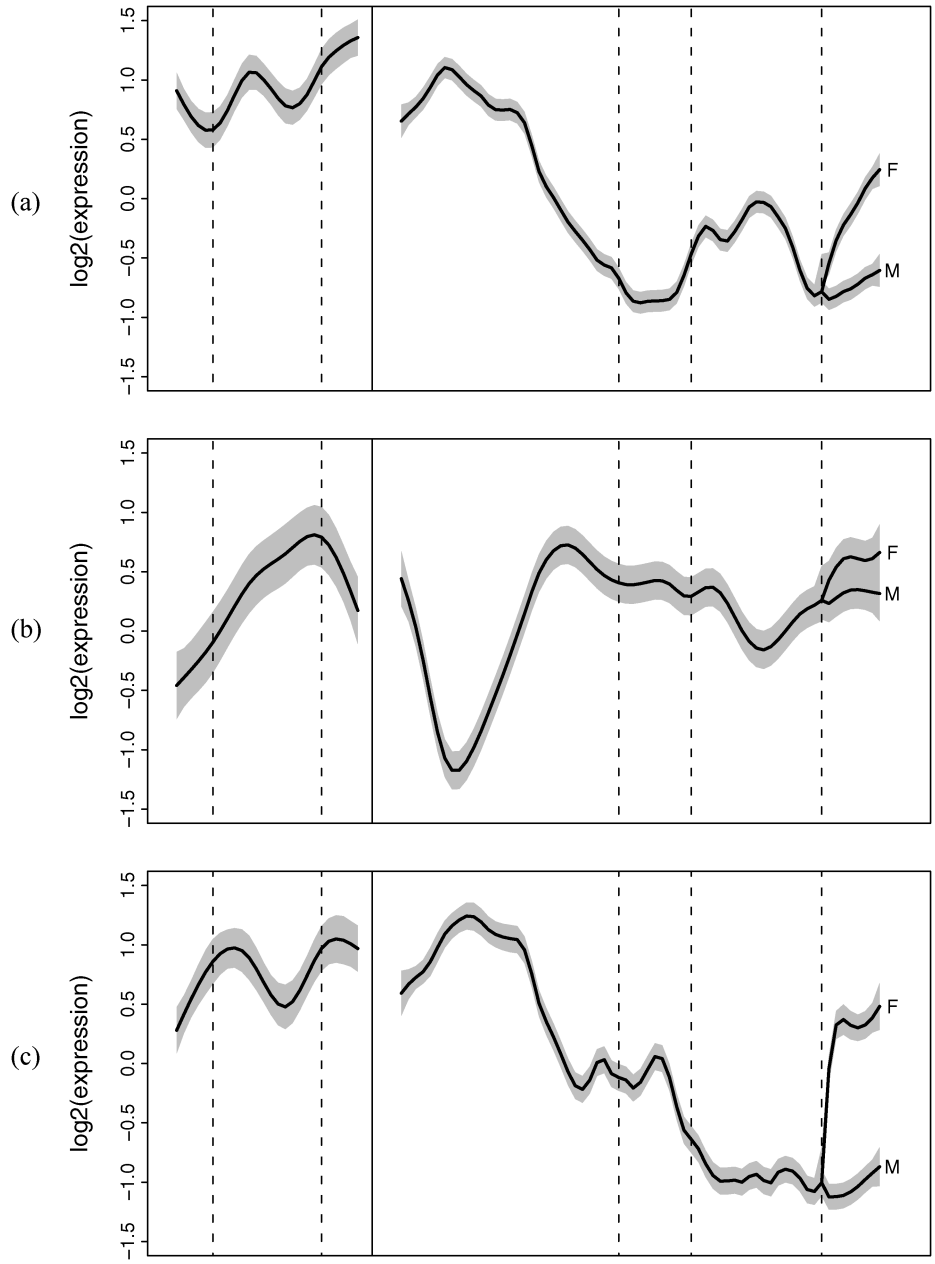

Figure 3. Estimated mean expression curves and 95% Bayesian confidence intervals (gray bands) for clusters A (a), B (b), and C (c) discovered in the worm-fly temporal expression data. Vertical solid lines separate worms (eggs, larva, and young adult are represented separately by dashed lines in the left frame) and fruitflies (embryogenesis, larva, pupa, and adult stages are separated by dash lines in the right frame). Adult fruitfly male and female mean expression curves are labeled M and F.

10 time points under anaerobic conditions. A reference sample pooled from all time points was used for hybridization.

For their analysis, Lai et al. (2006) normalized gene expressions to gene expressions of time 0 and filtered out differentially expressed genes. We used the normalized expressions at 23 time points of 2,388 differentially expressed genes for our clustering analysis. We modeled normalized gene expression $\mathbf{y}_i$ of the ith gene using the mixture model,

$$\mathbf{y}_i = \mu_k(\mathbf{t}, \tau) + b_i\mathbf{1} + \boldsymbol{\epsilon}_i \quad \text{with probability } p_k,$$

where $k = 1, \ldots, K$, $\tau = 1$ for aerobic conditions and $\tau = 2$ for anaerobic conditions, and $b_i \sim \mathrm{N}(0, \sigma_b^2)$ is the gene-specific random effect. We fit the model using the penalty (5) with $a = 2$. In total, 2,388 genes were clustered into 28 clusters using our method. We used FunSpec (Robinson, Grigull, Mohammad, and Hughes 2002) for gene annotation and biological function enrichment analysis. We found 26 out of 28 clusters with overrepresented biological functions. The estimated mean gene expression profiles and associated Bayesian confidence intervals of three clusters are given in Figure 4.

In cluster A, which consists of 57 genes, the estimated mean expression dropped progressively as oxygen level dropped, suggesting that the genes in this cluster were transiently downregulated in response to anaerobiosis. Furthermore, the estimated mean expression increased as oxygen concentration shifted back to the normal level. Accordingly, genes involved

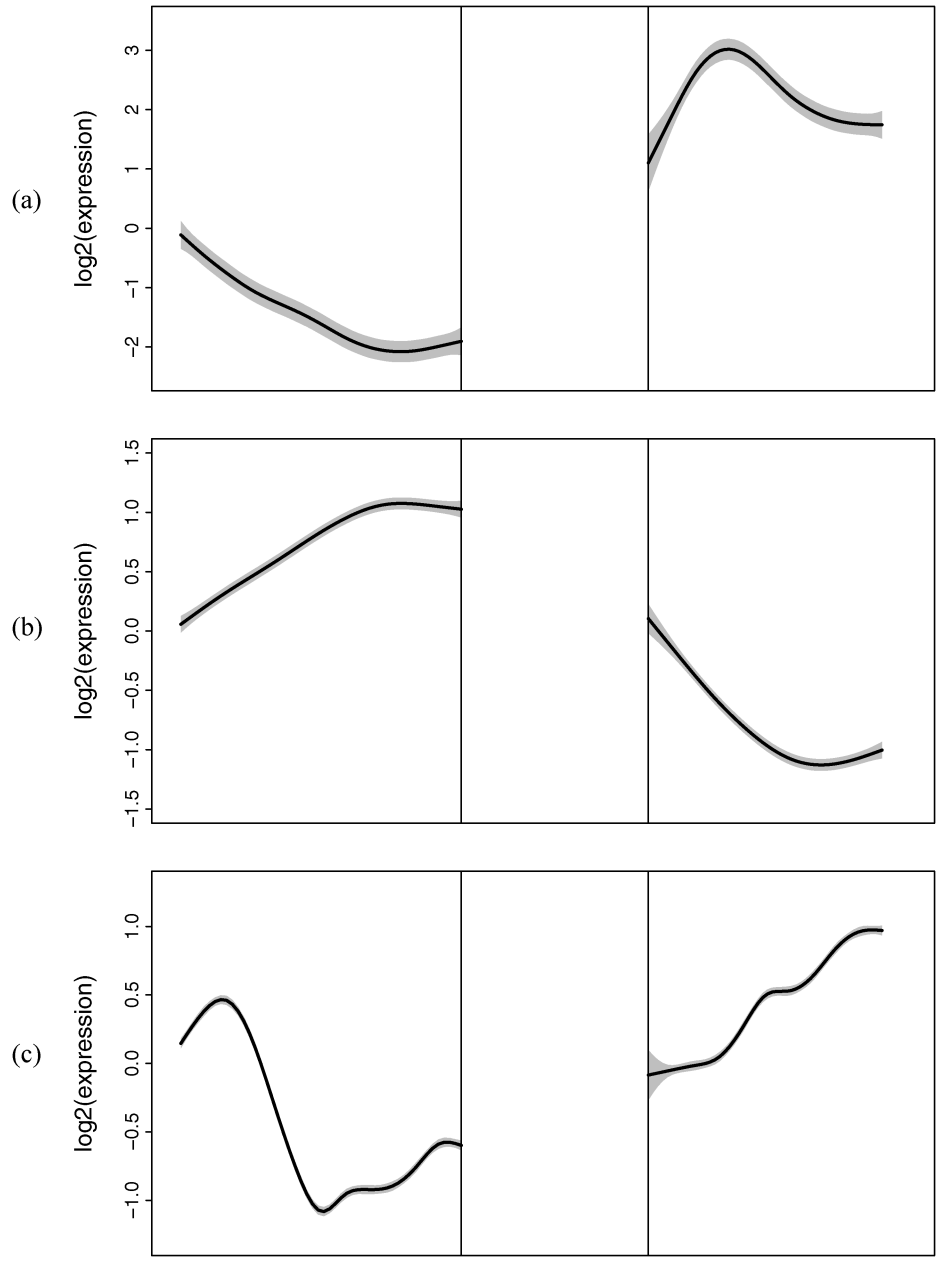

Figure 4. Estimated mean expression curves and 95% Bayesian confidence intervals (gray bands) for cluster A (a), B (b), and C (c) discovered in the yeast aerobic and anaerobic expression data. The aerobic (left) and anaerobic (right) conditions are separated by two vertical lines.

in respiration, lipid fatty acid and isoprenoid biosynthesis, and cell defense were overrepresented in this cluster ($p \leq 10^{-5}$).

In contrast to cluster A, cluster B (85 genes) consisted of genes involved in various biosyntheses, metabolism, and catabolism, such as glucose metabolism ($p \leq 10^{-6}$). These biological processes are necessary to maintain the basic living needs of yeast cells. Interestingly, the alcohol biosynthesis and metabolism also were enriched in this cluster. Consistent with biological function overrepresentation, the estimated mean expression was up-regulated in aerobic conditions and down-regulated in anaerobic conditions.

We had 70 genes in cluster C, where the estimated mean gene expression increased at the beginning and then dropped rapidly under aerobic conditions. Under anaerobic conditions, the estimated mean gene expression was up-regulated. In this cluster, respiratory deficiency and carbon utilization also were overrepresented ($p \leq 10^{-8}$). The initial up-regulation of gene expression under aerobic conditions can be explained in part by the fact that the cell increases energy up-take through other biological processes, such as carbon utilization, when oxygen decreases. But as the oxygen level continues to drop, these processes are replaced by more energy-efficient processes, such as glucose metabolism. Under the anaerobic conditions, these processes are revitalized as oxygen level increases.

6. DISCUSSION

In this article we have proposed a clustering method for large-scale functional data with multiple covariates. We have built nonparametric mixed-effects models that were nested under a mixture model. We used the penalized Henderson's like-

lihood for estimation and data-driven smoothing parameters, selected through GCV validation, to automatically capture the functional features. The rejection-controlled EM algorithm was designed to reduce the computational cost for large-scale data. The simulation analyses suggest that the proposed method outperforms the existing clustering methods. Moreover, the Bayesian interpretation of the proposed method allows the development of an equivalent fully Bayesian functional data clustering method that can accommodate additional genomic and proteomic information for gene expression studies. Although it was motivated for clustering temporal expression data, our proposed method has a wide spectrum of applications, including those involving seismic wave data arising from geophysical research (Wang, de Hoop, van der Hilst, Ma, and Tenorio 2006; Ma, Wang, Tenorio, de Hoop, and van der Hilst 2007). The calculations reported in this article were performed in R; open-source code is available in the R package MFDA.

As a sequel to this work, a clustering method for discrete data, especially those arising from temporal text mining, is under active development.

APPENDIX: PROOFS

Proof of Theorem 1

Note that if we specify the prior for f_0 as a Gaussian process with mean 0 and covariances $E[f_0(s_k)f_0(s_l)] = \tau^2 \sum_{\nu=1}^{m} \phi_\nu(s_k)\phi_\nu(s_l)$, then when $\tau^2 \to \infty$, the prior for f_0 becomes a diffuse prior (see Wahba 1983; Gu 2002). Assuming that $f_0(t)$ has a Gaussian process prior specified earlier, $f_1(x)$ has a Gaussian process prior specified as in Theorem 1, and $\mathbf{b}$ follows a normal distribution with mean 0 and variance–covariance matrix $\mathbf{B}$, we can derive that the joint distribution of $\mathbf{y}$ and $f_0(x) + f_1(x) + \mathbf{z}^T\mathbf{b}$ follows a Gaussian distribution with mean 0 and covariance matrix

$$\begin{pmatrix} b\mathbf{F}\mathbf{V}^+\mathbf{F}^T + \tau^2\mathbf{S}\mathbf{S}^T + \sigma^2\mathbf{I} & b\mathbf{F}\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \tau^2\mathbf{S}\boldsymbol{\phi} \\ b\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T + \tau^2\boldsymbol{\phi}^T\mathbf{S}^T & b\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \tau^2\boldsymbol{\phi}^T\boldsymbol{\phi} \end{pmatrix}, \tag{A.1}$$

where $\tilde{\boldsymbol{\xi}} = (R_1(s_1, x), \ldots, R_1(s_T, x), \mathbf{z})^T$ is $(T + p) \times 1$, $\boldsymbol{\phi}$ is $m \times 1$ with the νth entry $\phi_\nu(t)$, $\mathbf{F} = (\mathbf{R}, \mathbf{Z})$, and V^+ is the Moore–Penrose inverse of $\mathbf{V} = \operatorname{diag}(\mathbf{Q}, \frac{1}{N\lambda}\boldsymbol{\Omega})$ satisfying $\mathbf{V}\mathbf{V}^+\mathbf{F}^T = \mathbf{F}^T$.

Standard calculation yields

$$\begin{aligned} &E[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}] \\ &\quad = (b\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T + \tau^2\boldsymbol{\phi}^T\mathbf{S}^T)(b\mathbf{F}\mathbf{V}^+\mathbf{F}^T + \tau^2 SS^T + \sigma^2\mathbf{I})^{-1}\mathbf{y} \\ &\quad = \rho\boldsymbol{\phi}^T\mathbf{S}^T(\mathbf{W} + \rho\mathbf{S}\mathbf{S}^T)^{-1}\mathbf{y} + \tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T(\mathbf{W} + \rho\mathbf{S}\mathbf{S}^T)^{-1}\mathbf{y}, \end{aligned}$$

where $\rho = \tau^2/b$, $N\lambda = \sigma^2/b$, and $\mathbf{W} = \mathbf{F}\mathbf{V}^+\mathbf{F}^T + N\lambda\mathbf{I}$. Now, letting $\rho \to \infty$, we have

$$\lim_{\rho\to\infty}(\rho\mathbf{S}\mathbf{S}^T + \mathbf{W})^{-1} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1} \tag{A.2}$$

and

$$\lim_{\rho\to\infty}\rho\mathbf{S}^T(\rho\mathbf{S}\mathbf{S}^T + \mathbf{W})^{-1} = (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}. \tag{A.3}$$

(See Wahba 1983 and Gu 2002 for the proof.) Therefore, $\lim_{\tau^2\to\infty} E[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}] = \boldsymbol{\phi}^T\mathbf{d} + \tilde{\boldsymbol{\xi}}^T\tilde{\mathbf{c}}$, where

$$\begin{aligned} \mathbf{d} &= (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\mathbf{y}, \\ \tilde{\mathbf{c}} &= \mathbf{V}^+\mathbf{F}^T\big(\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\big)\mathbf{y}. \end{aligned} \tag{A.4}$$

It is straightforward to verify that the $\mathbf{d}$ and $\tilde{\mathbf{c}}$ given in (A.4) satisfy (8).

Proof of Theorem 2

The posterior variance can be easily calculated using expression (A.1) as follows:

$$\begin{aligned} \operatorname{var}[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}] = \tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \rho\boldsymbol{\phi}^T\boldsymbol{\phi} - (\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T + \rho\boldsymbol{\phi}^T\mathbf{S}^T) \\ \times (\mathbf{W} + \rho\mathbf{S}\mathbf{S}^T)^{-1}(\mathbf{F}\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \rho\mathbf{S}\boldsymbol{\phi}). \end{aligned}$$

Note that $\lim_{\rho\to\infty} \rho I - \rho^2\mathbf{S}^T(\rho\mathbf{S}\mathbf{S}^T + \mathbf{W})^{-1}\mathbf{S} = (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}$ and $\mathbf{V}\mathbf{V}^+\mathbf{F}^T = \mathbf{F}^T$. Therefore, as $\rho \to \infty$, we have

$$\begin{aligned} &\lim_{\tau^2\to\infty} \operatorname{var}[\mu(x) + \mathbf{z}^T\mathbf{b}|\mathbf{y}]/b \\ &\quad = \tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\boldsymbol{\phi} \\ &\qquad - 2\boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\mathbf{F}\mathbf{V}^+\tilde{\boldsymbol{\xi}} \\ &\qquad - \tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T\big(\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\big)\mathbf{F}\mathbf{V}^+\tilde{\boldsymbol{\xi}}. \end{aligned}$$

[Received June 2007. Revised January 2008.]

REFERENCES

Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M., Scott, M. P., Davis, R. W., and White, K. P. (2002), "Gene Expression During the Life Cycle of *Drosophila melanogaster*," *Science*, 297, 2270–2275.

Castillo-Davis, C., and Hartl, D. (2003), "Genemerge: Post-Genomic Analysis, Data-Mining and Hypothesis," *Bioinformatics*, 19, 891–892.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–37.

Dennis, J. E., and Schnabel, R. B. (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (corrected reprint of the 1983 original), Philadelphia: SIAM.

Fraley, C., and Raftery, A. E. (1990), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Green, P. J. (1990), "On the Use of the EM Algorithm for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society*, Ser. B, 52, 443–452.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer-Verlag.

Gu, C., and Ma, P. (2005), "Optimal Smoothing in Nonparametric Mixed Effect Models," *The Annals of Statistics*, 33, 1357–1379.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman & Hall.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006), "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves," *Journal of the American Statistical Association*, 101, 18–29.

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

James, G. M., and Sugar, C. A. (2003), "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association*, 98, 397–408.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S. K. (2001), "Genome-Wide Analysis of Developmental and Sex-Regulated Gene Expression Profiles in *Caenorhabditis elegans*," *Proceedings of the National Academy of Sciences*, 98, 218–223.

Kim, Y.-J., and Gu, C. (2004), "Smoothing Spline Gaussian Regression: More Scalable Computation via Efficient Approximation," *Journal of the Royal Statistical Society*, Ser. B, 66, 337–356.

Lai, L. C., Kosorukoff, A. L., Burke, P., and Kwast, K. E. (2006), "Metabolic State–Dependent Remodeling of the Transcriptome in Response to Anoxia and Subsequent Reoxygenation in *Saccharomyces cerevisiae*," *Eukayotic Cell*, 5, 1468–1489.

Liu, J. S., Chen, R., and Wong, W. H. (1998), "Rejection Control and Sequential Importance Sampling," *Journal of the American Statistical Association*, 93, 1022–1031.

Luan, Y., and Li, H. (2003), "Clustering of Time-Course Gene Expression Data Using a Mixed-Effects Models With B–Spline," *Bioinformatics*, 19, 474–282.

——— (2004), "Model-Based Methods for Identifying Periodically Regulated Genes Based on the Time Course Microarray Gene Expression Data," *Bioinformatics*, 20, 332–339.

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006), "A Data-Driven Clustering Method for Time Course Gene Expression Data," *Nucleic Acids Research*, 34, 1261–1269.

Ma, P., Wang, P., Tenorio, L., de Hoop, M. V., and van der Hilst, R. D. (2007), "Imaging of Structure at and Near the Core Mantle Boundary Using a Generalized Radon Transform: 2. Statistical Inference of Singularities," *Journal of Geophysical Research*, 112, B08303.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C. S., Jan, Y. N., Kenyon, C., Bargmann, C. I., and Li, H. (2004), "Comparing Genomic Expression Patterns Across Species Identifies Shared Transcriptional Program in Aging," *Nature Genetics*, 36, 197–204.

McLachlan, G. J., and Peel, D. (2001), *Finite Mixture Models*, New York: Wiley.

Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143.

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, New York: Springer-Verlag.

——— (2005), *Functional Data Analysis*, New York: Springer-Verlag.

Robinson, G. K. (1991), "That BLUP Is a Good Thing: The Estimation of the Random Effects" (with discussion), *Statistical Science*, 6, 15–51.

Robinson, M. D., Grigull, J., Mohammad, N., and Hughes, T. R. (2002), "Funspec: A Web-Based Cluster Interpreter for Yeast," *BMC Bioinformatics*, 3, 3–35.

Silverman, B. W., and Wood, J. T. (1987), "The Nonparametric Estimation of Branching Curves," *Journal of the American Statistical Association*, 82, 551–558.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, B., and Futcher, D. (1998), "Comprehensive Identification of Cell Cycle–Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization," *Molecular Biology of the Cell*, 9, 3273–3297.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R., and Davis, R. G. (2005), "Significance of Time Course Microarray Experiments," *Proceedings of the National Academy of Sciences*, 102, 12837–12842.

Wahba, G. (1983), "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Ser. B, 45, 133–150.

——— (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

Wang, P., de Hoop, M. V., van der Hilst, R. D., Ma, P., and Tenorio, L. (2006), "Imaging of Structure at and Near the Core Mantle Boundary Using a Generalized Radon Transform: 1. Construction of Image Gathers," *Journal of Geophysical Research*, 111, B12304.

Wang, Y. (1998), "Mixed-Effects Smoothing Spline ANOVA," *Journal of the Royal Statistical Society*, Ser. B, 60, 159–174.

Wei, G. C., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.