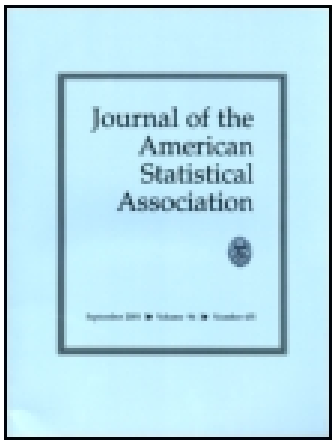


Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories With Application to Cocaine Abuse Treatment Data

Hui Huang, Yehua Li & Yongtao Guan

Accepted author version posted online: 01 Oct 2014.Published online: 22 Dec 2014.

To cite this article: Hui Huang, Yehua Li & Yongtao Guan (2014) Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories With Application to Cocaine Abuse Treatment Data, Journal of the American Statistical Association, 109:508, 1412-1424, DOI: 10.1080/01621459.2014.957286

To link to this article: http://dx.doi.org/10.1080/01621459.2014.957286

Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories With Application to Cocaine Abuse Treatment Data

Hui HUANG, Yehua LI, and Yongtao GUAN

In a cocaine dependence treatment study, we have paired binary longitudinal trajectories that record the cocaine use patterns of each patient before and after a treatment. To better understand the drug-using behaviors among the patients, we propose a general framework based on functional data analysis to jointly model and cluster these paired non-Gaussian longitudinal trajectories. Our approach assumes that the response variables follow distributions from the exponential family, with the canonical parameters determined by some latent Gaussian processes. To reduce the dimensionality of the latent processes, we express them by a truncated Karhunen-Lóeve (KL) expansion allowing the mean and covariance functions to be different across clusters. We further represent the mean and eigenfunctions functions by flexible spline bases, and determine the orders of the truncated KL expansions using data-driven methods. By treating the cluster membership as a missing value, we cluster the cocaine use trajectories by a likelihood-based approach. The cluster membership and parameter estimates are jointly estimated by a Monte Carlo EM algorithm with Gibbs sampling steps. We discover subgroups of patients with distinct behaviors in terms of overall probability to use, binge verses periodic use pattern, etc. The joint modeling approach also sheds new lights on relating relapse behavior to baseline pattern in each subgroup. Supplementary materials for this article are available online.

KEY WORDS: Clustering; EM algorithm; Exponential family; Functional data analysis; Joint modeling; Metropolis-Hastings algorithm; Splines.

1. INTRODUCTION

1.1 Data Description

In this article, we consider data in the form of longitudinal trajectories that are commonly collected in cocaine and other substance abuse treatment studies. The data that we will analyze came from a recently completed clinical trial by the Yale Stress Center. In the study, 142 cocaine dependent subjects were admitted to the Clinical Neuroscience Research Unit of the Connecticut Mental Health Center, for two to four weeks of in-patient relapse prevention treatment for cocaine dependence. Upon treatment entry, all subjects recalled their daily cocaine use retrospectively for the previous ninety days by using the time-line follow-back (TLFB; Sobell and Sobell 1993) Substance Use Calendar, which is a reliable and well established instrument for assessing self-report drug use in alcoholic and drug abusing populations (Fals-Stewart et al. 2000). After completion of the in-patient treatment, all participants were invited back for follow-up interview(s) to assess cocaine use outcomes. The study was conducted in two phases. For the 59 subjects who participated the first phase of the study, only one interview was administrated at day 90 after the treatment. For the remaining eighty three subjects enrolled in the second phase of the study, four interviews were given at days 14, 30, 90, and 180 after the treatment. Daily cocaine use records were recalled based on the TLFB procedure during each interview for the period prior to the interview date.

Even though the actual daily cocaine use amounts (in gram equivalents) were estimated by the study participants, such data are often subject to large errors because of the long period of time associated with these recalls as well as the lack of a common scale to assess the amount used due to different methods of consumption (e.g., intranasal use versus injection). We therefore consider only dichotomized trajectories comprised of no use (=0) and any use (=1). We refer to the resulting trajectories before and after the treatment as the baseline and relapse trajectories, respectively. Figure 1 shows the baseline and relapse trajectories from three subjects. Clearly different subjects exhibit very different cocaine use patterns in terms of use frequencies. In particular, subject 3 also appears to be exhibit a strong weekly cycle.

In a cocaine abuse treatment study like ours, an important goal is to understand the causes for the often highly diverse treatment outcomes, which have been given in the forms of daily cocaine use trajectories in our case. It is of particular interest to know the relationship between one's baseline cocaine use pattern and the treatment outcome. For example, do high frequency users during the baseline period tend to use more after treatment, and will those with a weekly baseline cocaine use pattern maintain such a pattern after the treatment? If a significant relationship can be found, then we can potentially provide a much improved prediction for one's treatment prognosis based on his/her baseline cocaine use pattern. More significantly, such a knowledge could even help us design more desirable treatment plans, should multiple treatment options be available. We answer the questions raised above by jointly modeling and clustering the paired baseline and relapse trajectories. The clustering analysis will allow

Hui Huang is Research Fellow, Center for Statistical Science and School of Mathematical Science, Peking University, Beijing, China, 100871 (E-mail: huanghui@math.pku.edu.cn). Yehua Li is Associate Professor, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: yehuali@iastate.edu). Yongtao Guan is Professor, Department of Management Science, University of Miami, Coral Gables, FL 33124 (E-mail: yguan@bus.miami.edu). This research has been partially supported by NIH grant 7R01DA029081 and NSF grants DMS-0845368, DMS-1105634, and DMS-1317118.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

DOI: 10.1080/01621459.2014.957286

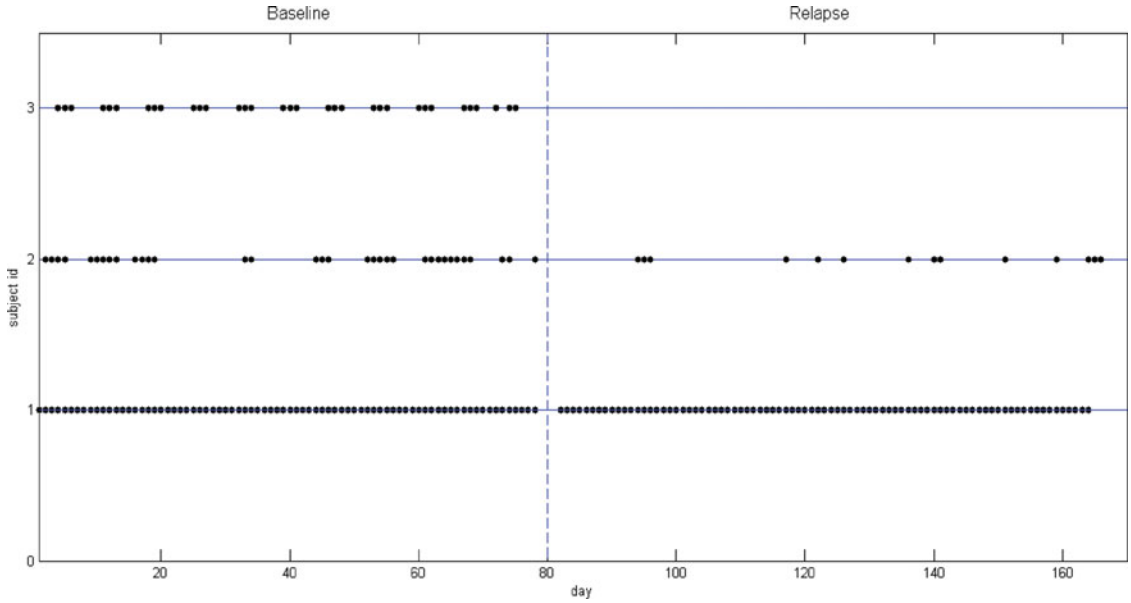


Figure 1. Illustration of individual cocaine use. The vertical dotted line in day 80 separate the whole observations to baseline (left) and relapse (right) trajectories. Subject 1 (bottom line) uses cocaine almost everyday; Subject 2 (middle line) has an irregular cocaine use trajectory; Subject 3 (top line) has a weekly pattern before the treatment and no use after the treatment.

us to assess the different types of cocaine use behaviors both before and after a treatment, and the joint modeling can provide insight on how one's baseline cocaine use pattern may affect the treatment outcome. The mean and covariance structures of these trajectories can be highly complex (e.g., Guan, Li, and Sinha 2011) and therefore be difficult to model by a parametric approach. We will instead develop a more flexible, nonparametric approach based on functional data analysis (Ramsay and Silverman 2005).

1.2 Literature Review

In functional data analysis, functional principal component analysis (FPCA) is a useful tool to capture features of both mean and covariance functions. There is a vast volume of recent literature on FPCA, including James, Hastie, and Sugar (2000); Yao, Müller, and Wang (2005a); Hall, Müller, and Wang (2006); and Li and Hsing (2010). These papers studied FPCA for a single functional variable. More recently, Di et al. (2009) and Crainiceanu, Staicu, and Di (2009) considered multilevel FPCA, while Zhou et al. (2010) studied hierarchical functional data. In all three papers, the random curves are assumed to be spanned by the same set of eigenfunctions. In contrast, Yao, Müller, and Wang (2005b) and Zhou, Huang, and Carroll (2008) modeled pairs of functional variables, where the two functional variables are allowed to have different sets of eigenfunctions and their principal component scores are cross-correlated. Hall, Müller, and Yao (2008) extended the FPCA to non-Gaussian longitudinal data. They modeled the longitudinal response by a generalized linear mixed model with a known link function, where the latent longitudinal process is assumed to have a FPCA decomposition.

There is also a strong interest in clustering functional data. James and Sugar (2003) proposed a model-based clustering method for sparsely sampled functional data. Jiang and Serban (2011) considered clustering spatially correlated functional data. Both papers assumed that the random curves had different mean curves across clusters but the same covariance structure. Serban and Jiang (2012) studied clustering multilevel functional data under a modeling framework similar to that in Di et al. (2009). Chiou and Li (2007, 2008) proposed to cluster functional data using the k-means method. There are also some related methods for clustering non-Gaussian data in the machine learning literature, such as the model-based methods (Rao and Scott 1992; Banfield and Raftery 1993), the K-means and its variants (Ordonez 2003), and the entropy-based clustering methods (Li 2006). These methods, however, were not designed for longitudinal data and hence not suitable for our problem.

Our proposed method advances the existing literature in two innovative ways. First, joint modeling and clustering of non-Gaussian trajectories are unified in one framework. A joint modeling approach is needed for our motivating data, because one's baseline cocaine use pattern may be related to relapse. To achieve this, we use a generalized linear mixed model to link the binary daily cocaine use records to a latent longitudinal process that admits an FPCA decomposition. At the latent process level, we use a joint modeling approach as in Zhou, Huang, and Carroll (2008) to model the paired longitudinal latent processes, allowing them to have different eigen-systems. The FPCA approach in our joint model can reduce the dimensionality of the latent functional data efficiently, and increase the interpretability of the model. As James and Sugar (2003), we adopt a model-based clustering framework, by modeling the class labels as multinomial random variables. To handle computational challenges due to non-Gaussian responses, we propose a Monte Carlo EM (MCEM) algorithm based on Metropolis-Hasting steps for parameter estimation. Second, we allow the latent processes to have different mean and different covariance structures across clusters, where in contrast most existing literature only allows the former. In cocaine abuse treatment studies, it is also important to study the covariance structure, because it may carry useful information about one's cocaine use behavior. For

example, a covariance structure with a long dependence range often suggests a binge pattern, while that exhibiting weekly cycles indicates a weekly cocaine use pattern. Even though the model will be more complex by allowing different covariance structures, this modeling effort is not formidable because we reduce the dimensionality of the data efficiently by FPCA. We organize the remainder of this article as follows. We present our proposed model in Section 2, and describe its estimation through MCEM in Section 3. We discuss issues on model selection in Section 4. We present some simulation results in Section 5, and apply the proposed method to our motivating data in Section 6. We discuss the strengths and limitations of our approach in Section 7. Technical details are in the Appendices, and more details of the simulation study can be found in the supplementary material.

2. STATISTICAL MODELING

2.1 A General Modeling Framework

The data are collected from n independent subjects. Let $\mathcal{T}$ and $\mathcal{S}$ be the time intervals of the baseline and followup study period, and $\{B_i(t), t \in \mathcal{T}\}$ and $\{R_i(s), s \in \mathcal{S}\}$ denote the cocaine use records for the ith subject in the baseline and followup period respectively. The observed generalized responses are $B_{ij} = B_i(t_{ij})$, $j = 1, \ldots, T_i$, and $R_{i\ell} = R_i(s_{i\ell})$, $\ell = 1, \ldots, S_i$. In our motivating data, $B_i(t)$ and $R_i(s)$ are both binary trajectories, but we will develop the proposed method in a more general setting.

We assume that, conditional on some latent random factors, the distributions of B_{ij} and $R_{i\ell}$ belong to the canonical exponential family with densities

$$\begin{aligned} f(B_{ij}|\gamma_{B,ij}, \sigma_1) &= \exp\left[\frac{B_{ij}\gamma_{B,ij} - b_1(\gamma_{B,ij})}{a_1(\sigma_1)} + c_1(B_{ij}, \sigma_1)\right], \\ & j = 1, 2, \ldots, T_i, \\ f(R_{i\ell}|\gamma_{R,i\ell}, \sigma_2) &= \exp\left[\frac{R_{i\ell}\gamma_{R,i\ell} - b_2(\gamma_{R,i\ell})}{a_2(\sigma_2)} + c_2(R_{i\ell}, \sigma_2)\right], \\ & \ell = 1, 2, \ldots, S_i, \end{aligned} \tag{2.1}$$

where $\gamma_{B,ij}$ and $\gamma_{R,i\ell}$ are canonical parameters depending on some latent random factors, σ_k, $k = 1, 2$, are dispersion parameters, and $a_k(\cdot)$, $b_k(\cdot)$, and $c_k(\cdot)$, $k = 1, 2$, are known functions (McCulloch and Nelder 1989). We assume

$$\gamma_{B,ij} = X_i(t_{ij}), \quad \gamma_{R,i\ell} = Y_i(s_{i\ell}),$$

where $\{X_i(t), t \in \mathcal{T}\}$ and $\{Y_i(s), s \in \mathcal{S}\}$ are two latent Gaussian processes that drive the behavior of the observed longitudinal processes. In the cocaine abuse treatment study, $X_i(t)$ and $Y_i(s)$ can be understood as unobserved temporally varying factors such as cocaine craving levels that may affect one's probability of cocaine use on a given day. Conditioning on $X_i(t)$ and $Y_i(s)$, we assume that $B_{i1}, \ldots, B_{iT_i}, R_{i1}, \ldots, R_{iS_i}$ are mutually independent.

We assume that the subjects are drawn from C populations, and all information in $B_i(t)$ and $R_i(s)$ related to the cluster membership is completely specified by the latent processes $[X_i(t), Y_i(t)]$. Define the latent cluster label for the ith subject as $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iC})'$ where

$$\omega_{ic} = \begin{cases} 1 & \text{if subject i belongs to cluster } c \\ 0 & \text{otherwise,} \end{cases}$$

for $c = 1, \ldots, C$. Assume that $\boldsymbol{\omega}_i \sim$ Multinomial$(1, [\pi_1, \ldots, \pi_C])$, where $\pi_c > 0$ for all $c = 1, \ldots, C$ and $\sum_{c=1}^{C} \pi_c = 1$.

Let $\mu_{x,c}(t) = \mathrm{E}[X_i(t)|\omega_{ic} = 1]$ and $\Gamma_{x,c}(t_1, t_2) = \mathrm{cov}[X_i(t_1), X_i(t_2)|\omega_{ic} = 1]$ be the mean and covariance function of $X(\cdot)$ in cluster c. Similarly, define $\mu_{y,c}(s)$ and $\Gamma_{y,c}(s_1, s_2)$ as the mean and covariance of $Y(\cdot)$ in cluster c. We allow $(\mu_{x,c}, \Gamma_{x,c}, \mu_{y,c}, \Gamma_{y,c})$ to vary across different clusters. We also allow different covariance structures across clusters to better understand the cocaine use pattern in different sub populations of cocaine users.

In our application, some of the baseline trajectories showed periodic patterns and we also anticipate certain trend in the relapse trajectories. We hence model $\mu_{x,c}$ and $\mu_{y,c}$ nonparametrically. Guan, Li, and Sinha (2011) examined the empirical autocorrelation of individual baseline trajectories. They found that the correlation varied across subjects, decayed at a rate much slower than low order autoregressive correlations and often exhibited periodic patterns. It is challenging to fully capture such a complex correlation structure parametrically. As a result, we model $\Gamma_{x,c}$ and $\Gamma_{y,c}$ flexibly as bivariate nonparametric functions. To keep the model computationally tractable, we propose to reduce the dimensionality of the covariance functions using the recently much-celebrated FPCA method.

Suppose that the covariance functions yield the spectral decomposition, $\Gamma_{x,c}(t_1, t_2) = \sum_{j=1}^{\infty} \lambda_{jx,c}\psi_{jc}(t_1)\psi_{jc}(t_2)$ and $\Gamma_{y,c}(s_1, s_2) = \sum_{j=1}^{\infty} \lambda_{jy,c}\phi_{jc}(s_1)\phi_{jc}(s_2)$, where $\lambda_{1x,c} \geq \lambda_{2x,c} \geq \cdots > 0$ and $\lambda_{1y,c} \geq \lambda_{2y,c} \geq \cdots > 0$ are the eigenvalues of $\Gamma_{x,c}(\cdot, \cdot)$ and $\Gamma_{y,c}(\cdot, \cdot)$ respectively, and $\psi_{jc}(\cdot)$ and $\phi_{jc}(\cdot)$ are the corresponding eigenfunctions. The eigenfunctions are orthonormal functions so that $\int_{\mathcal{T}} \psi_{jc}(t)\psi_{j'c}(t)dt = I(j = j')$ and $\int_{\mathcal{S}} \phi_{jc}(s)\phi_{j'c}(s)ds = I(j = j')$ for all j, j' and c, where $I(\cdot)$ is an indicator function. Given the cluster membership of subject i, we can then express $X_i(t)$ and $Y_i(s)$ using the Karhunen-Loève expansion

$$\begin{aligned} \text{for } \omega_{ic} = 1, \quad X_i(t) &= \mu_{x,c}(t) + \sum_{j=1}^{\infty} \xi_{ij}\psi_{jc}(t), \\ Y_i(s) &= \mu_{y,c}(s) + \sum_{j=1}^{\infty} \zeta_{ij}\phi_{jc}(s), \end{aligned} \tag{2}$$

where ξ_{ij} and ζ_{ij} are normal random variables with mean zero and variances equal to $\lambda_{jx,c}$ and $\lambda_{jy,c}$. In practice, model (2) is not feasible for estimation due to the infinite dimensionality of the parameter space. Instead, it is common to approximate (2) by using the leading principal components, that is,

$$\begin{aligned} X_i(t) &\approx \mu_{x,c}(t) + \sum_{j=1}^{P_{1,c}} \xi_{ij}\psi_{jc}(t), \\ Y_i(s) &\approx \mu_{y,c}(s) + \sum_{j=1}^{P_{2,c}} \zeta_{ij}\phi_{jc}(s), \end{aligned} \tag{3}$$

where $P_{1,c}$ and $P_{2,c}$ are sufficiently large to capture most variation in the random processes. The reduced rank model (3) can help obtain more reliable estimates, especially for irregular and/or sparse functional data (James, Hastie, and Sugar 2000; Zhou, Huang, and Carroll 2008). We treat $P_{1,c}$ and $P_{2,c}$ as

tuning parameters and choose them using data driven methods in Section 4.

The Karhunen-Loève theorem ensures that ξ_{ij}'s are uncorrelated between different orders of j, and so are the variables ζ_{ij}'s. Let $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ be two column vectors containing all ξ_{ij}'s and ζ_{ij}'s used in the reduced rank model (3). As discussed in Section 1, it is of interest to understand how one's baseline cocaine-use behavior is related to his/her relapse pattern. In our joint modeling framework, we therefore allow $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ to be cross-correlated. Following Zhou, Huang, and Carroll (2008) and Yao, Müller, and Wang (2005b), we jointly model $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ by

$$\begin{pmatrix} \boldsymbol{\xi}_i \\ \boldsymbol{\zeta}_i \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D}_{\lambda \mathbf{x},\mathbf{c}} & \boldsymbol{\Sigma}_{xy,c} \\ \boldsymbol{\Sigma}_{yx,c} & \mathbf{D}_{\lambda y,c} \end{pmatrix} \right], \tag{4}$$

where $\mathbf{D}_{\lambda x,c} = \text{diag}\{\lambda_{1x,c}, \lambda_{2x,c}, \ldots, \lambda_{P_{1,c}x,c}\}$, $\mathbf{D}_{\lambda y,c} = \text{diag}\{\lambda_{1y,c}, \lambda_{2y,c}, \ldots, \lambda_{P_{2,c}y,c}\}$, $\boldsymbol{\Sigma}_{xy,c}$ is a $P_{1,c} \times P_{2,c}$ cross-covariance matrix, and $\boldsymbol{\Sigma}_{yx,c} = \boldsymbol{\Sigma}_{xy,c}^T$. To better quantify the association between $\boldsymbol{\xi}_i$ and ζ_i, we also define the cross-correlation matrix as

$$\mathcal{R}_{xy,c} = D_{\lambda x,c}^{-1/2} \boldsymbol{\Sigma}_{xy,c} D_{\lambda y,c}^{-1/2} = \{\rho_{jk,c}\}, \tag{5}$$

where $\rho_{jk,c} = \text{corr}(\xi_{ij}, \zeta_{ik} | \omega_{ic} = 1)$ for $j = 1, \ldots, P_{1,c}$ and $k = 1, \ldots, P_{2,c}$.

Our conditional modeling of the generalized longitudinal trajectories in (1) resembles that in Hall, Müller, and Yao (2008), but there are three major differences between our approach and theirs. First, the procedure proposed by Hall, Müller, and Yao (2008) is designed for modeling a single non-Gaussian longitudinal process but not for jointly modeling paired longitudinal processes. Second, even though they adopted a conditional modeling of the longitudinal process, their estimation method is a marginal regression method based on asymptotic approximations under the assumption that the latent Gaussian process has a "small" magnitude. Third, even though the estimated principal component scores produced by Hall et al. can be further used for clustering in a two-stage procedure, the estimation error in the first stage will inevitably propagate into the clustering analysis. In contrast, our model based approach treats estimation and clustering under a unified framework and seems to be a more natural approach for this problem.

2.2 Reduced Rank Spline Representation

Assuming that the mean functions and eigenfunctions in model (3) are smooth, we approximate them by reduced rank regression splines. Specifically, let $\mathbf{S}_1(t)$ and $\mathbf{S}_2(s)$ be q_1 and q_2 dimensional B-spline bases defined on $\mathcal{T}$ and $\mathcal{S}$ respectively. For simplicity, we place the knots of the spline bases to be equally spaced in $\mathcal{T}$ and $\mathcal{S}$. Using the normalization method in Zhou, Huang, and Carroll (2008), we orthogonalize the spline bases so that

$$\int \mathbf{S}_1(t)\mathbf{S}_1^T(t)dt = I_{q_1 \times q_1}, \quad \int \mathbf{S}_2(s)\mathbf{S}_2^T(s)ds = I_{q_2 \times q_2},$$

where $\boldsymbol{I}_{m \times m}$ is an m-dimension identity matrix. We then express the mean functions and eigenfunctions as

$$\begin{aligned} \mu_{x,c}(t) &= \mathbf{S}_1^T(t)\boldsymbol{\nu}_{x,c}, \quad [\psi_{1c}(t), \ldots, \psi_{P_{1,c}}(t)] = \mathbf{S}_1^T(t)\boldsymbol{\Theta}_{\psi,c}, \\ \mu_{y,c}(s) &= \mathbf{S}_2^T(s)\boldsymbol{\nu}_{y,c}, \quad [\phi_{1c}(s), \ldots, \phi_{P_{2,c}}(s)] = \mathbf{S}_2^T(s)\boldsymbol{\Theta}_{\phi,c}, \end{aligned} \tag{6}$$

where $\boldsymbol{\nu}_{x,c}, \boldsymbol{\nu}_{y,c}, \boldsymbol{\Theta}_{\psi,c}$, and $\boldsymbol{\Theta}_{\phi,c}$ are coefficients with dimensions $q_1 \times 1$, $q_2 \times 1$, $q_1 \times P_{1,c}$, and $q_2 \times P_{2,c}$, respectively. Based on Lemma 1 in Zhou, Huang, and Carroll (2008), the identifiability of $\boldsymbol{\Theta}$, $\mathbf{D}$ and $\boldsymbol{\Sigma}_{xy}$ is guaranteed by two conditions: orthornormality and consistent signs of eigenfunctions. Thus, we set the following constraints on $\boldsymbol{\Theta}_{\psi,c}$ and $\boldsymbol{\Theta}_{\phi,c}$:

$$\boldsymbol{\Theta}_{\psi,c}^T \boldsymbol{\Theta}_{\psi,c} = I_{P_{1,c} \times P_{1,c}}, \quad \boldsymbol{\Theta}_{\phi,c}^T \boldsymbol{\Theta}_{\phi,c} = I_{P_{2,c} \times P_{2,c}}. \tag{7}$$

In addition, for each eigenfunction, we force the sign of the first element with the largest magnitude to be positive. The reduced rank model (3) can now be expressed as

$$\begin{aligned} X_i(t) &= \mathbf{S}_1^T(t)\boldsymbol{\nu}_{x,c} + \mathbf{S}_1^T(t)\boldsymbol{\Theta}_{\psi,c}\boldsymbol{\xi}_i, \\ Y_i(s) &= \mathbf{S}_2^T(s)\boldsymbol{\nu}_{y,c} + \mathbf{S}_2^T(s)\boldsymbol{\Theta}_{\phi,c}\boldsymbol{\zeta}_i. \end{aligned} \tag{8}$$

With the spline approximation described above, the parameters to be estimated are collected in $\boldsymbol{\Omega} = \{\boldsymbol{\Omega}_c, c = 1, \ldots, C\}$, where $\boldsymbol{\Omega}_c = \{\pi_c, \boldsymbol{\nu}_{x,c}, \boldsymbol{\nu}_{y,c}, \boldsymbol{\Theta}_{\psi,c}, \boldsymbol{\Theta}_{\phi,c}, \mathbf{D}_{\lambda x,c}, \mathbf{D}_{\lambda y,c}, \boldsymbol{\Sigma}_{xy,c}\}$.

Let $\mathbf{B}_i$ and $\mathbf{R}_i$ be the data vectors collecting all observations in the baseline and relapse trajectories for subject i, then his/her contribution to the complete data likelihood is

$$L_i(\boldsymbol{\Omega}; \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\omega}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i) = \prod_{c=1}^{C} \left\{ \pi_c f_c(\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i) \left[\prod_{j=1}^{T_i} f_c(B_{ij} | \boldsymbol{\xi}_i) \right] \left[\prod_{\ell=1}^{S_i} f_c(R_{i\ell} | \boldsymbol{\zeta}_i) \right] \right\}^{\omega_{ic}}, \tag{9}$$

where $f_c(\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i)$ is the joint density of the latent variables specified in (4), and $f_c(B_{ij}|\boldsymbol{\xi}_i)$ and $f_c(R_{i\ell}|\boldsymbol{\zeta}_i)$ are the conditional distributions specified in model (1) given that the ith subject belongs to cluster c. In the cocaine abuse treatment data, both $B_i(t)$ and $R_i(s)$ are binary processes and their conditional distributions are given by

$$\begin{aligned} f_c(B_{ij}|\boldsymbol{\xi}_i) &= [\alpha_{i,x}(t_{ij})]^{B_{ij}} \cdot [1 - \alpha_{i,x}(t_{ij})]^{1-B_{ij}}, \\ f_c(R_{i\ell}|\boldsymbol{\zeta}_i) &= [\alpha_{i,y}(s_{i\ell})]^{R_{i\ell}} \cdot [1 - \alpha_{i,y}(s_{i\ell})]^{1-R_{i\ell}}, \end{aligned}$$

where $\alpha_{i,x}(t) = g[\mathbf{S}_1^T(t)\boldsymbol{\nu}_{x,c} + \mathbf{S}_1^T(t)\boldsymbol{\Theta}_{\psi,c}\boldsymbol{\xi}_i]$, $\alpha_{i,y}(s) = g[\mathbf{S}_2^T(s)\boldsymbol{\nu}_{y,c} + \mathbf{S}_2^T(s)\boldsymbol{\Theta}_{\phi,c}\boldsymbol{\zeta}_i]$, and $g(x) = \exp(x)/[1 + \exp(x)]$.

To facilitate the Gibbs sampler described in the Appendix, it is more convenient to rewrite (4) in a regression form (Zhou, Huang, and Carroll 2008)

$$\boldsymbol{\xi}_i = \boldsymbol{\Lambda}_c \cdot \boldsymbol{\zeta}_i + \boldsymbol{\eta}_{ic}, \tag{10}$$

where $\boldsymbol{\Lambda}_c = \boldsymbol{\Sigma}_{xy,c}\mathbf{D}_{\lambda y,c}^{-1}$ and $\boldsymbol{\eta}_{ic}$ is a zero-mean normal vector independent with $\boldsymbol{\zeta}_i$ and with a covariance matrix $\boldsymbol{\Sigma}_{\eta,c} = \mathbf{D}_{\lambda x,c} - \boldsymbol{\Sigma}_{xy,c}\mathbf{D}_{\lambda y,c}^{-1}\boldsymbol{\Sigma}_{yx,c}$. Then, the conditional joint density of $(\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i)$ can be written as

$$\begin{aligned} f_c(\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i) &= f_c(\boldsymbol{\xi}_i|\boldsymbol{\zeta}_i) f_c(\boldsymbol{\zeta}_i) \\ &= \frac{1}{|\boldsymbol{\Sigma}_{\eta,c}|^{1/2} |\mathbf{D}_{\lambda y,c}|^{1/2}} \\ &\quad \exp\left[-\frac{1}{2} (\boldsymbol{\xi}_i - \boldsymbol{\Lambda}_c \cdot \boldsymbol{\zeta}_i)^T \boldsymbol{\Sigma}_{\eta,c}^{-1} (\boldsymbol{\xi}_i - \boldsymbol{\Lambda}_c \cdot \boldsymbol{\zeta}_i) \right] \exp\left(-\frac{1}{2} \boldsymbol{\zeta}_i^T \mathbf{D}_{\lambda y,c}^{-1} \boldsymbol{\zeta}_i \right). \end{aligned}$$

3. ESTIMATION PROCEDURE

3.1 Monte Carlo EM Algorithm

The complete data likelihood (9) depends on the latent random variables $\boldsymbol{\omega}_i$, $\boldsymbol{\xi}_i$, and $\boldsymbol{\zeta}_i$ and hence cannot be maximized

directly. We treat these latent variables as missing data, and estimate the unknown parameters by a Monte Carlo EM algorithm (Wei and Tanner 1990; McCulloch 1997), which was shown by Chan and Ledolter (1995) to converge to the maximum likelihood estimator under some general regularity conditions. Cluster memberships can be determined by assigning subject i to cluster c that maximizes the conditional probability $\pi_{ic} = P(\omega_{ic} = 1|\mathbf{B}_i, \mathbf{R}_i)$. The procedure iterates between an E-step and an M-step described in the following subsections until the convergence of the parameters.

3.1.1 E-Step. In the E-step, we take expectation on the logarithm of the likelihood (9) conditioning on the observed data $\mathbf{B}_i, \mathbf{R}_i$ and the value of the parameter vector $\boldsymbol{\Omega}_{\text{prev}}$ from the previous EM iteration. For non-Gaussian responses, this conditional expectation does not have a closed form, hence we approximate it by a Monte Carlo average.

Specifically, let $\ell_i(\boldsymbol{\Omega}; \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\omega}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i)$ be the log complete data likelihood for the ith subject given in (9). In the E-step, we need to evaluate the expected log-likelihood

$$Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) = \sum_{i=1}^{n} \mathrm{E}[\ell_i(\boldsymbol{\Omega}; \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\omega}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i)|\mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}}].$$

We approximate $Q(\cdot)$ by its Monte Carlo counterpart,

$$\widehat{Q}(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \ell_i \left(\boldsymbol{\Omega}; \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\omega}_i^{(k)}, \boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}\right), \quad (11)$$

where K is the Monte Carlo sample size and $\boldsymbol{\omega}_i^{(k)}$, $\boldsymbol{\xi}_i^{(k)}$, $\boldsymbol{\zeta}_i^{(k)}$ are samples from the conditional distribution of $[\boldsymbol{\omega}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i|\mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}}]$.

Since this conditional distribution does not have a closed form, we draw samples from it using the Gibbs sampler (Geman and Geman 1984) incorporated with a Metropolis-Hastings step (Tanner 1993). The detailed algorithm is provided in Appendix A.

3.1.2 M-Step. In the M-step, we update the parameters to their current values, $\boldsymbol{\Omega}_{\text{curr}}$, which maximize $\widehat{Q}(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}})$ defined in (11). Up to some negligible constant terms, we factorize $\widehat{Q}(\cdot)$ into

$$\begin{aligned}
\widehat{Q}(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) = {} & \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{c=1}^{C} \omega_{ic}^{(k)} \log(\pi_c) \\
& + \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{c=1}^{C} \omega_{ic}^{(k)} \sum_{j=1}^{T_i} \log\left[f_c\left(B_{ij}|\boldsymbol{\xi}_i^{(k)}\right)\right] \\
& + \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{c=1}^{C} \omega_{ic}^{(k)} \sum_{\ell=1}^{S_i} \log\left[f_c\left(R_{i\ell}|\boldsymbol{\zeta}_i^{(k)}\right)\right] \\
& + \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{c=1}^{C} \omega_{ic}^{(k)} \left\{\log\left[f_c\left(\boldsymbol{\xi}_i^{(k)}|\boldsymbol{\zeta}_i^{(k)}\right)\right]\right. \\
& \left. + \log\left[f_c\left(\boldsymbol{\zeta}_i^{(k)}\right)\right]\right\} \\
\doteq {} & \widehat{Q}_1(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) + \widehat{Q}_2(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) + \widehat{Q}_3(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) \\
& + \widehat{Q}_4(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}). \quad (12)
\end{aligned}$$

One can easily see that $\widehat{Q}_1$, $\widehat{Q}_2$, $\widehat{Q}_3$, and $\widehat{Q}_4$ depend on mutually disjoint sets of parameters in $\boldsymbol{\Omega}$ and therefore can be maximized separately. Specifically, π_c's are updated by maximizing $\widehat{Q}_1$, $(\boldsymbol{\nu}_{x,c}, \boldsymbol{\Theta}_{\psi,c})$ and $(\boldsymbol{\nu}_{y,c}, \boldsymbol{\Theta}_{\phi,c})$ are updated by maximizing $\widehat{Q}_2$ and $\widehat{Q}_3$, and finally $(\boldsymbol{\Lambda}_c, \boldsymbol{\Sigma}_{\eta,c}, D_{\lambda_{yc}})$ are updated by maximizing $\widehat{Q}_4$. When maximizing $\widehat{Q}_2$ and $\widehat{Q}_3$, we apply iterative reweighted least square (IRLS) method and update $\boldsymbol{\Theta}_{\psi,c}$ and $\boldsymbol{\Theta}_{\phi,c}$ under the constraint (7). The detailed algorithm for the M-step is provided in Appendix B.

3.2 Implementation and Cluster Membership Estimation

We stop the EM iterations when the parameters satisfy

$$\max_l \left(\frac{|\Omega_{l,\text{curr}} - \Omega_{l,\text{prev}}|}{|\Omega_{l,\text{prev}}| + \delta_1}\right) < \delta_2, \quad (13)$$

where δ_1 and δ_2 are predetermined constants, Ω_l denotes the lth parameter in $\boldsymbol{\Omega}$. Following the suggestion in Booth and Hobert (1999), we set $\delta_1 = 0.001$ and $\delta_2 = 0.005$.

Increasing the Monte Carlo sample size K can reduce the Monte Carlo errors in (11), but may waste computation time especially during the early iterations when the parameter values are still far from the truth. As a compromise, we gradually increase K along the iterations. Specifically, in our simulation study and data analysis, we set $K = 500$ for iterations 1–4, $K = 5000$ for iterations 5–9, $K = 10{,}000$ for iterations 10–19 and $K = 20{,}000$ for iterations 20 and over.

At convergence of the algorithm, we estimate the conditional probability of subject i belonging to cluster c given the observed data as $\widehat{\pi}_{ic} = K^{-1} \sum_{k=1}^{K} \omega_{ic}^{(k)}, c = 1, \ldots, C$, where $\omega_{ic}^{(k)}$'s are the Monte Carlo samples at the final E-step. We assign the cluster membership of subject i to $\arg\max_c\{\widehat{\pi}_{ic}\}$.

4. MODEL SELECTION

4.1 Information Criterion

Model selection is one of the most important issues in model-based clustering problems. In our setting, there are three key features that need to be determined, that is, the number of clusters C, the numbers of leading principal components $\{P_{1,c}\}_{c=1}^{C}$ and $\{P_{2,c}\}_{c=1}^{C}$, and the dimensions q_1 and q_2 of the spline bases $\mathbf{S}_1$ and $\mathbf{S}_2$ given in Equation (8). Let $\mathcal{M}$ be a candidate model, and $N_{\mathcal{M}}$ be the effective number of parameters in $\mathcal{M}$, which is defined as the total number of parameters in $\mathcal{M}$ minus the number of constraints in (7). We choose the model that minimizes the following Baysian information criterion (BIC):

$$\text{BIC}(\mathcal{M}) = -2\widehat{Q}_{\mathcal{M}} + N_{\mathcal{M}} \cdot \log(n), \quad (4.14)$$

where n is the number of subjects and $\widehat{Q}_{\mathcal{M}}$ is the Monte Carlo average defined in (11) using the Monte Carlo samples from the final EM iteration.

The BIC in (14) is a special case of the IC_Q criterion in Ibrahim, Zhu, and Tang (2008). Instead of using the log-likelihood of the observed data, which does not have a close form, they used the Monte Carlo expectation of the logarithm of the complete data likelihood. By using the final outputs of the MCEM algorithm, it does not require any extra effort to evaluate the IC_Q criterion. Ibrahim, Zhu, and Tang (2008) performed intensive simulation studies on an AIC version of (14) which replaces the penalty term with $2N_{\mathcal{M}}$, and it performed well in various missing data settings. They also proposed a more

sophisticated $IC_{H,Q}$ criterion, but that procedure is considerably more difficult to implement in our setting. We use the BIC version of IC_Q to impose higher penalty on overestimating the model, and this procedure performs well in our simulation study.

4.2 Expedited Search for the Optimal Model

The MCEM algorithm is computationally intensive, and it therefore can be extremely time consuming to search over all possible models in the model selection procedure. We take further steps to expedite model selection by narrowing the search range.

Among the features that determine model complexity, the numbers of spline basis functions q_1 and q_2 are of least importance. The inverse of the number of spline bases is asymptotically equivalent to the bandwidth in a kernel smoother, and the spline estimators are consistent for a wide range of values for q_1 and q_2. Assuming the mean and eigenfunctions in our model are twice continuously differentiable, we follow the asymptotic theory of Li and Hsing (2010) to set $q_1 \approx (nT)^{1/5} + 4$ and $q_2 \approx (nS)^{1/5} + 4$, where T and S be averages of T_i's and S_i's. One can fix q_1 and q_2 at these suggested values or do a quick search in a neighborhood of these values.

In contrast to the basis functions, the number of clusters C and the numbers of principal components in each cluster, that is, $\{P_{1,c}\}_{c=1}^{C}$ and $\{P_{2,c}\}_{c=1}^{C}$, are much more crucial. To further expedite the search, we assume that the number of principal components are the same across clusters, that is $P_1 \equiv P_{1,c}$ and $P_2 \equiv P_{2,c}$, for $c = 1, \ldots, C$. With these approximations, the BIC in (14) only depends on (C, P_1, P_2), and we can minimize the BIC using a three-dimension grid search. Further simplification can be made to narrow the search range for P_1, P_2, and C, and we present such a strategy in the supplementary material.

5. SIMULATION STUDY

We conduct a simulation study to assess the performance of our proposed clustering and estimation method. In the simulation, $n = 200$ subjects are randomly drawn from $C = 2$ clusters, with marginal probabilities $\pi_1 = 0.6$ from cluster 1 and $\pi_2 = 0.4$ from cluster 2. The latent processes $X_i(t)$ and $Y_i(s)$ are spanned by $P_1 = 2$ and $P_2 = 1$ eigenfunctions respectively. We set the mean and eigenfunctions of these processes to be different in different clusters. Details on simulating the latent processes, including the eigenvalues, cross-correlation parameters between X_i and Y_i as well as the explicit mathematical expressions and plots of the mean and eigen functions are provided in the supplementary material. We first generate the latent processes, and then generate the binary baseline and relapse trajectories $B_i(t)$ and $R_i(s)$ based on model (1) with a logistic link. We repeat the simulation 100 times. For each simulated dataset, we use the method in Section 4 to determine (C, P_1, P_2). The true model with $(C, P_1, P_2) = (2, 2, 1)$ is correctly selected 95% of the time. This high accuracy provides a strong support for our model selection approach.

We study the performance of clustering using the adjusted Rand index. The Rand index measures the agreement between two partitions of objects (Rand 1971). Hubert and Arabie (1985) proposed an improved version of the index which has an expected value of 0 and is bounded by ± 1. A larger adjusted Rand index means a higher similarity between the two partitions. In our simulation, we calculate the adjusted Rand index between the clustering results and the true cluster memberships in each simulation, and then take the average over the 100 simulations to evaluate the effectiveness of our proposed method. For comparison, we perform two additional analyses. In the first analysis, we concatenate the baseline and relapse trajectories and apply the conventional k-means method, assuming that all subjects have the same covariance structure. In the second analysis, we perform clustering and estimation based on baseline and relapse trajectories alone. The average adjusted Rand Indices are 0.9376 for our proposed method, 0.7381 for the k-means method, and 0.7725 and 0.7441 by using baseline and relapse trajectories alone, respectively. These results indicate the superiority of our method over these existing alternative methods.

To assess the accuracy of the estimation, we calculate the mean square error (MSE) for the scalar parameters, and the mean integrated squared error (MISE) for the mean and eigen functions. The detailed results are given in Tables W.1–W.3 in the supplementary material. Compared to the approach using baseline or relapse trajectories alone, our proposed joint modeling approach leads to considerable reductions in MSEs (mean reduction: 44%, range: 8% to 73%) for the scalar parameters and moderate reductions in MISE (mean reduction: 12%, range: 3% to 22%) for the mean and eigen functions. Note that the cross-correlation parameters can only be estimated by the joint modeling approach.

We also provide some graphical summary of the estimated functions in Figures W.4 and W.5 in the supplementary material. The true $\mu(\cdot)$ functions, the mean and the 5th and 95th percentiles of the estimated curves based on the joint model are given in Figure W.4. Similar plots for $\phi(\cdot)$'s and $\psi(\cdot)$'s are provided in Figure W.5. From these plots, we find that the means of the estimated functions are almost identical to the true functions, indicating very small biases. Moreover, the percentile bands are all very narrow indicating small variances in these functional estimators.

6. COCAINE USE DATA ANALYSIS

We apply our proposed method to cluster the baseline and relapse trajectories of cocaine use described in Section 1. In light of the potential weekly use patterns in the baseline period, we align the baseline trajectories in such a way that all baseline trajectories started on the first Monday of the baseline period and lasted for 80 days. For the relapse trajectories, we will only consider data in the first 90 days, because there were significantly more missing data in the second 90-day period. In particular, the 59 subjects enrolled in the first phase of the study had no data reported 90 days after the treatment. We use the actual number of post-treatment days to index the relapse trajectories. Conventional approaches to analyze similar data are typically conducted under a regression framework. Specifically, summary statistics extracted from the relapse trajectories such as percent days abstinent and time to first relapse are often used as dependent variables, whereas summary statistics derived from the baseline trajectories such as frequency and average daily use amount are included as predictors (e.g., Carroll et al. 1993; Anton et al. 2006; Sinha et al. 2006).

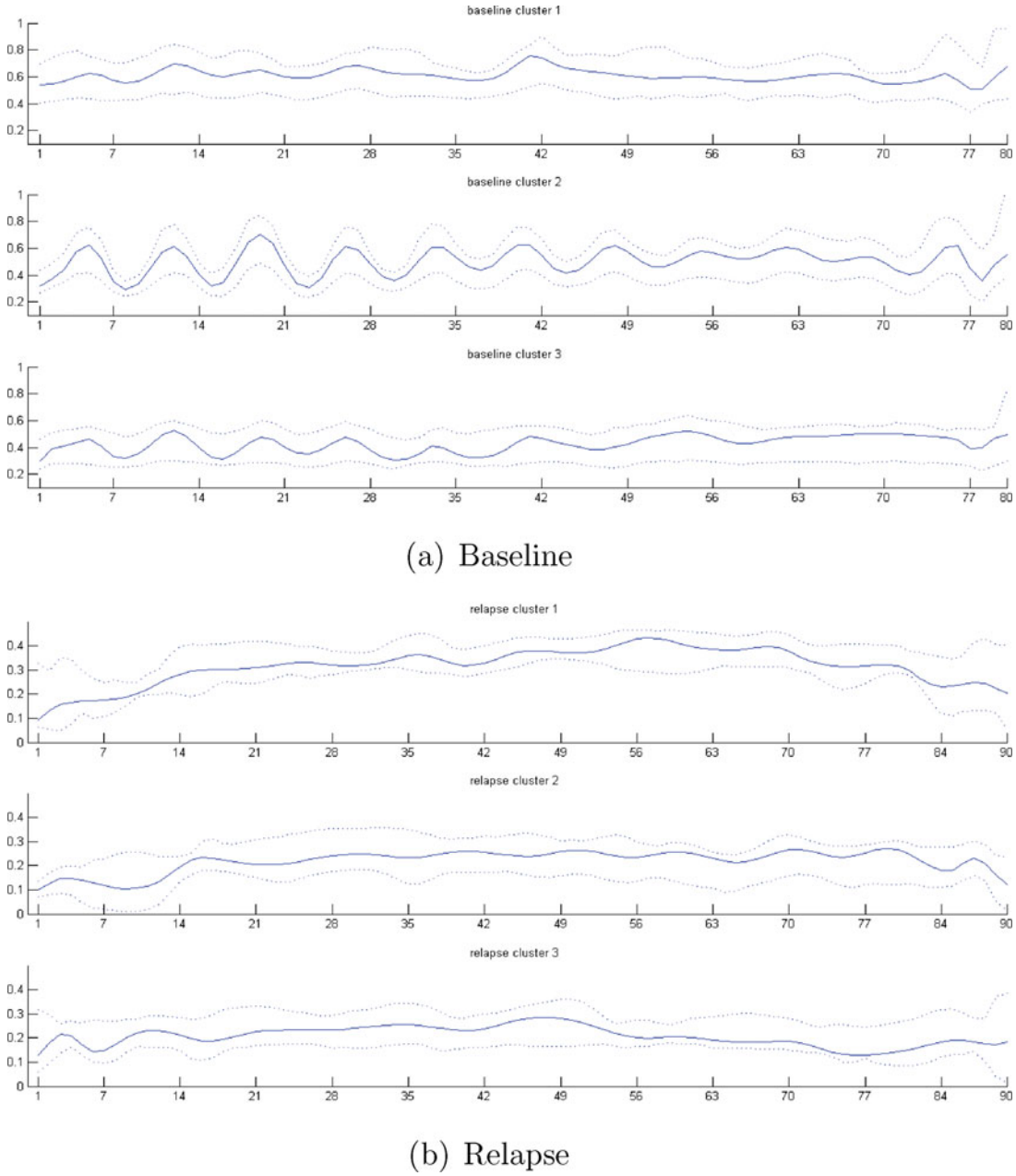


Figure 2. Estimated probabilities (solid lines) and their 95% bootstrap CI (dotted lines) for each cluster.

For modeling the mean and eigenfunctions, we use cubic B-spline bases with seven interior knots, where the interior knots are equally spaced over the baseline and relapse periods, respectively. We determine the Monte Carlo size and the stopping rule as described in Section 3.2 and closely monitor the convergence of algorithm. Additional trace plots for the Gibbs sampler and the EM iterations are provided in the supplementary material.

Following the procedure described in Section 4, we obtain the minimum BIC score from the model with $C = 3$, $P_1 = 3$, and $P_2 = 2$. In other words, we obtain three clusters, and in each cluster there are three and two leading FPCs for the baseline and relapse trajectories, respectively. At the convergence of MCEM, the numbers of subjects in each cluster are $n_1 = 41$, $n_2 = 59$, and $n_3 = 33$. To summarize the three clusters, we find that the average daily use frequency for baseline and relapse periods are 0.73 and 0.17 in cluster 1, 0.43 and 0.14 for cluster 2, and 0.37 and 0.08 for cluster 3. Our estimation results using the proposed model provide further insights about the mean and covariance structures of the trajectories in each cluster than these simple summary statistics.

6.1 Results on Mean Structures of the Trajectories

Figure 2 shows the estimated mean probabilities of daily cocaine use, together with the associated pointwise 95% bootstrap confidence intervals based on 100 bootstrap replicates, where the bootstrap procedure is carried out by resampling subjects within each cluster. The estimated mean probabilities describe the mean structures of the trajectories in different clusters. For all clusters, the probabilities for the baseline trajectories are always higher than those for the relapse trajectories, which indicates the

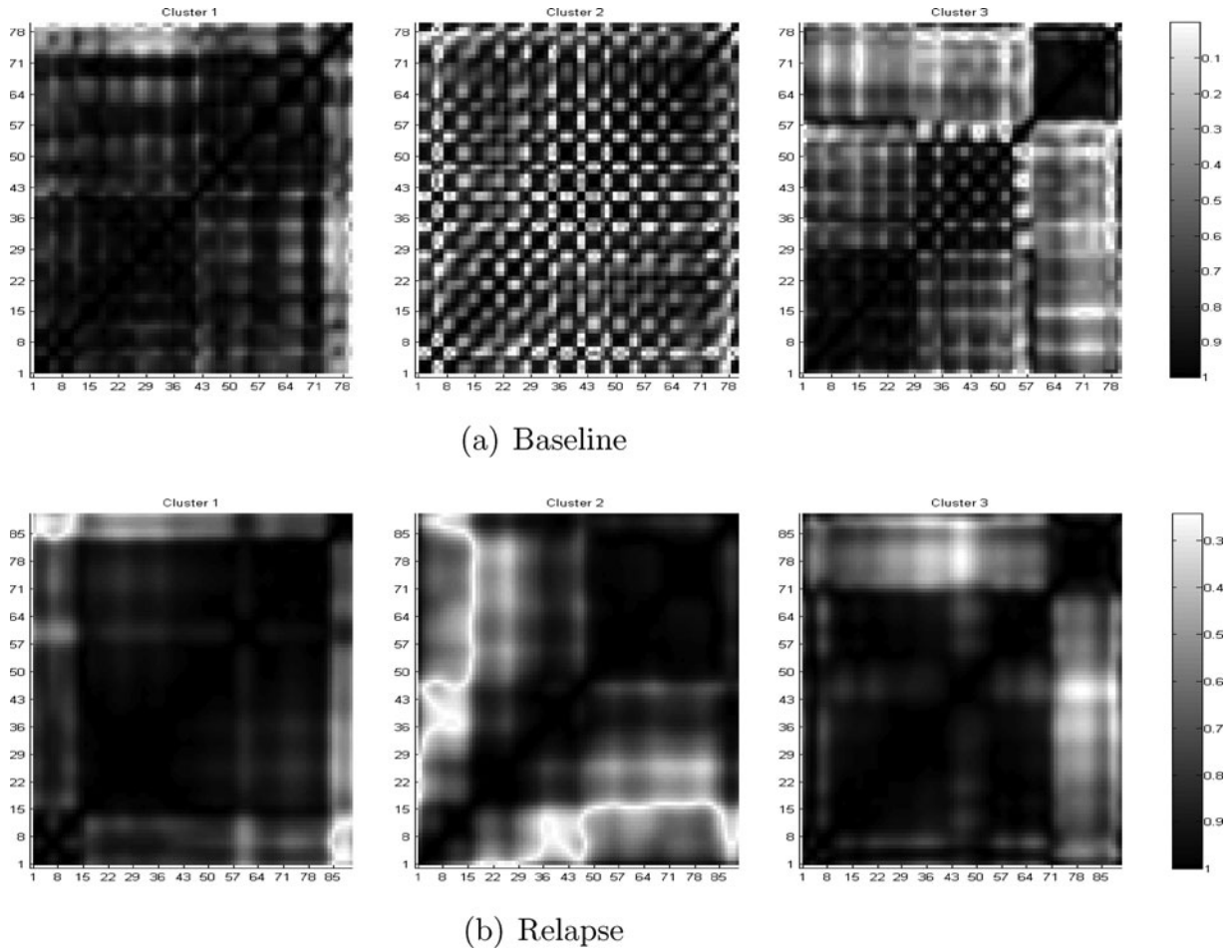


Figure 3. Estimated correlation surfaces. Darker level means higher correlation.

benefit of treatment. For the baseline trajectories, cluster 1 has overall the highest probabilities, followed in turn by clusters 2 and 3. The same order holds for the relapse trajectories. This observation suggests that one's baseline cocaine use pattern is related to relapse, in that a heavier cocaine user in the baseline period tends to use more cocaine after the treatment. Such a phenomenon has been well documented in literature (e.g., Fox et al. 2006).

There appear to be weekly patterns in the baseline period for the subjects in clusters 2 and 3, with the pattern being more prominent in cluster 2. The peaks generally occurred on Fridays, indicating that subjects in these two clusters were potentially recreational weekend users. Despite the local bumps in the estimated probabilities, the overall trend of the baseline trajectories appears to be flat. In contrast, the estimated probabilities for the relapse trajectories are smoother but show an increasing trend with time for both clusters 1 and 2, indicating that the treatment effect diminished over time. The increase is especially fast and significant for subjects in cluster 1, who were heavy cocaine users. It may appear puzzling at the first sight that the probabilities drop after day 60 for these subjects. However, the drop was likely due to increased dropout and incarceration rates, in which cases the subjects either failed to report any cocaine use or were forced to be abstinent. Hence, the treatment did not appear to be particularly effective for subjects in cluster 1.

6.2 Results on Correlation Structures of the Trajectories

Based on the estimated leading eigenvalues and FPCs, we estimate the covariance for the baseline and relapse trajectories as

$$\widehat{\Gamma}_{x,c}(t_1,t_2)=\sum_{j=1}^{3}\widehat{\lambda}_{jx,c}\widehat{\psi}_{jc}(t_1)\widehat{\psi}_{jc}(t_2),\quad t_1,t_2=1,\ldots,80,$$

$$\widehat{\Gamma}_{y,c}(s_1,s_2)=\sum_{j=1}^{2}\widehat{\lambda}_{jy,c}\widehat{\phi}_{jc}(s_1)\widehat{\phi}_{jc}(s_2),\quad s_1,s_2=1,\ldots,90.$$

Given the estimated covariance, we can further estimate the correlation and investigate how the correlation differs across clusters. The top panel of Figure 3 shows the estimated correlation surfaces for the baseline trajectories. There appears to be a periodic pattern in the plots for clusters 2 and 3, and a further look reveals that the cycle is seven days. This indicates a weekly baseline cocaine use pattern for subjects in these two clusters. However, no such weekly pattern is observed in the surface plot for cluster 1. The bottom panel of Figure 3 shows the estimated correlation surfaces for the relapse trajectories and none of these plots shows any weekly pattern. This suggests that the weekly cocaine users during the baseline period no longer maintained any weekly pattern after the treatment. Hence, the treatment effectively changed the cocaine use behaviors of subjects in these two clusters.

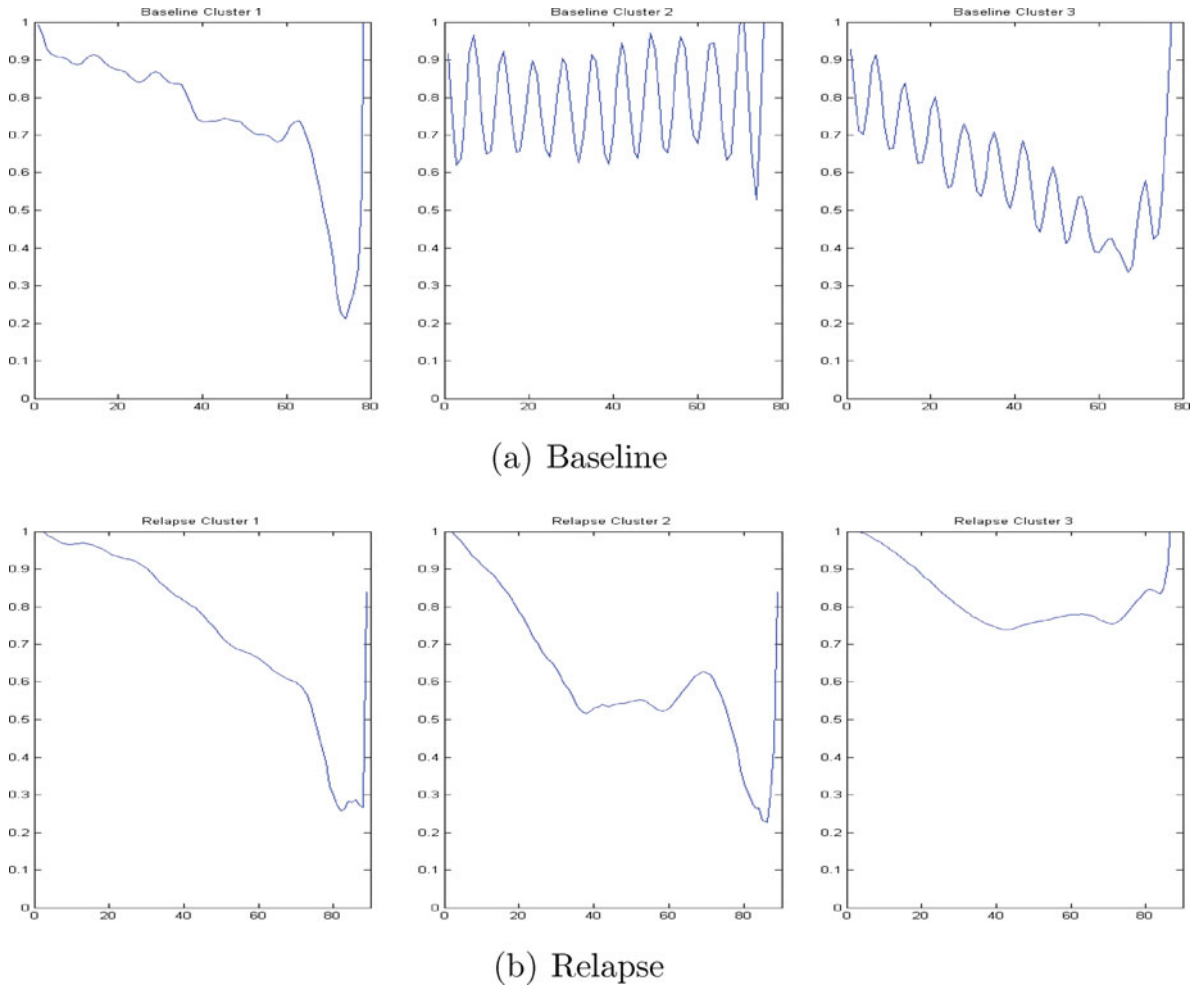


Figure 4. Averaged correlations versus time lags.

For a further investigation, we average the correlations as

$$\overline{\text{Corr}}_{xc}(r) = \frac{\sum_{|t_1-t_2|=r} \widehat{\text{Corr}}_{xc}(t_1, t_2)}{\sum_{|t_1-t_2|=r} 1},$$
$$\overline{\text{Corr}}_{yc}(r) = \frac{\sum_{|s_1-s_2|=r} \widehat{\text{Corr}}_{yc}(s_1, s_2)}{\sum_{|s_1-s_2|=r} 1},$$

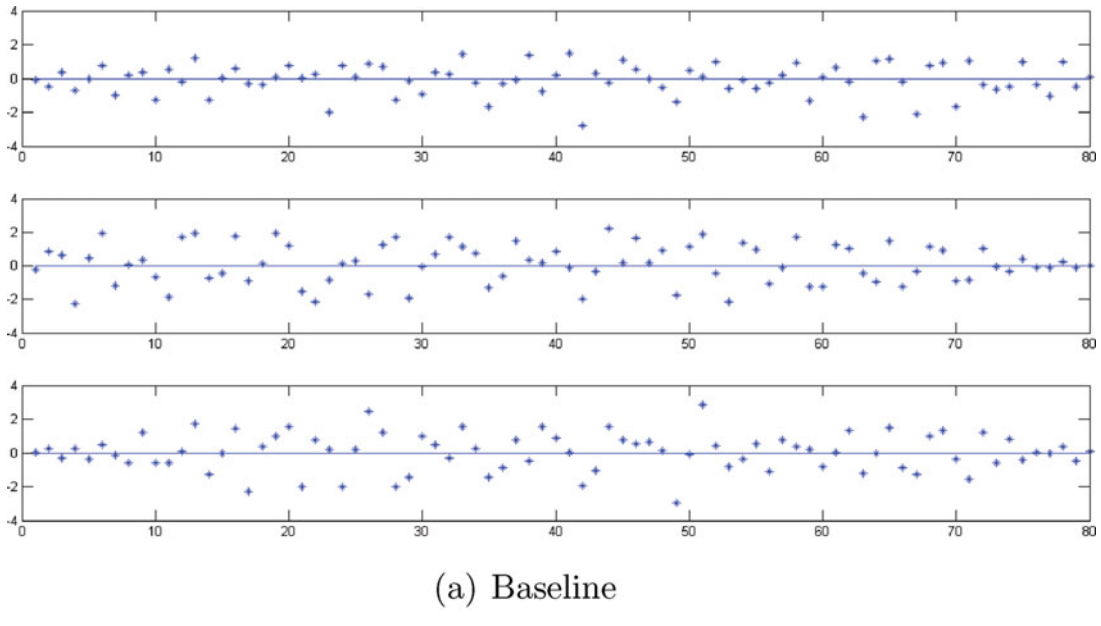
where r is the time lag. The top panel of Figure 4 shows plots of the average correlation for the baseline trajectories. Note that the weekly patterns become more obvious for both clusters 2 and 3. However, there is also a clear difference between them. Specifically, the average correlation tends to decrease as the time lag increases for cluster 3, but no such trend is present for cluster 2. A decreasing trend in the current context would suggest that cocaine use outcomes are less correlated when the time lag is larger. Hence, subjects in cluster 2 exhibited a more regular cocaine use pattern than those in cluster 3. We stress that such a conclusion cannot be reached by examining the mean structures alone. Note that a decreasing trend is also observed in the plot for cluster 1.

The bottom panel of Figure 4 shows plots of the average correlation for the relapse trajectories. The overall trend in these plots is decreasing for all three clusters. There are some irregular alterations at the last few time lags. For a larger time lag, the sample size used to produce the average correlation is smaller. This in turn will increase the variability of the estimate. The average correlation is large (> 0.7) at all r values for cluster 3. From Figure 2(b), we also see that the estimated probabilities of daily cocaine use are similar over time for the relapse trajectories in cluster 3. Combined together, these two results suggest that the relapse pattern of subjects in this cluster did not change significantly during the study period. In other words, the treatment appeared to have had a long-lasting and positive effect on cocaine users in this particular group.

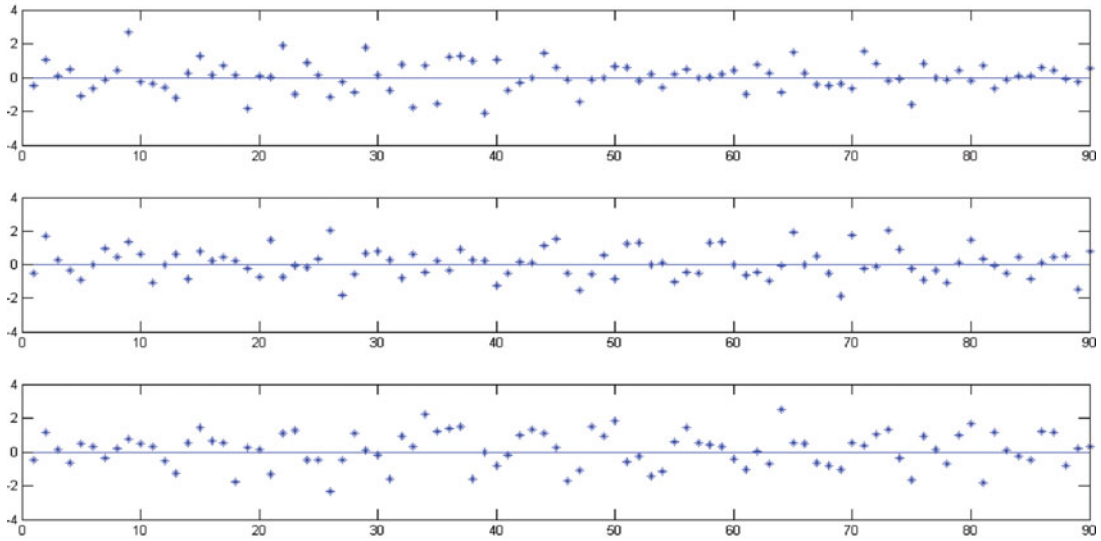
6.3 Results on Cross-Correlation Between Trajectories

Let $\widehat{\rho}_{jk,c}$ ($j = 1, 2, 3$, $k = 1, 2$, $c = 1, 2, 3$) denote the estimated correlation between the jth FPC for the baseline trajectories and the kth FPC for the relapse trajectories in cluster c. Table W.4 in the supplementary material shows the estimates of these cross-correlation parameters and their 90% bootstrap confidence intervals. These results show that most of these correlation parameters are significantly different from zero, which indicates strong correlation between the baseline and relapse behaviors of a patient.

It is of particular interest to study $\widehat{\rho}_{11,c}$, because it contributes the most to the correlation between the baseline and relapse trajectories. For the baseline trajectories, the first FPCs explain 75%, 91%, 82% of variation for cluster 1, 2, and 3, respectively. For the relapse trajectories, the percentages of variation



(a) Baseline



(b) Relapse

Figure 5. Plots of standardized differences between observed frequencies and expected probabilities of cocaine use over time at baseline and relapse within clusters.

explained by the first FPCs become 82%, 90%, and 83%. The values of $\widehat{\rho}_{11,c}$ are 0.43, 0.17, and 0.28 for clusters 1, 2, and 3, respectively, with 90% confidence intervals $(0.13, 0.75)$, $(0.07, 0.32)$, and $(0.12, 0.36)$. The positive values of $\widehat{\rho}_{11,c}$ suggest that within each cluster, if one used more cocaine in the baseline period, he/she tended to use more after the treatment. This result agrees with a similar finding that we made by visually comparing the estimated probabilities of daily cocaine use shown in Figure 2. The larger $\widehat{\rho}_{11,c}$ is, the stronger the relationship is. Hence, the treatment outcomes were the least affected by their baseline pattern for users in cluster 2 (i.e., the strong weekly cocaine users), followed in turn by those in clusters 3 and 1. This result reaffirms that the treatment was effective in altering weekly cocaine users' cocaine use pattern. It also warrants further investigations to understand the mechanism causing the change. From a practical point of view, our results indicate that it is important to treat weekly cocaine users differently from the rest such as binge users. This is significant because most existing studies in literature often do not single them out as one particular group of interest.

6.4 Model Diagnostics

To assess the goodness-of fit of the fitted model, we calculate differences between observed frequencies and expected probabilities of cocaine use over time at baseline and relapse within clusters. Given cluster membership c, we define the standardized difference for baseline at day t as

$$\varepsilon_{x,c}(t) = \frac{\sum_{i\in c}[B_i(t) - \hat{\alpha}_{i,x}(t)]}{\sqrt{\sum_{i\in c}\hat{\alpha}_{i,x}(t)[1-\hat{\alpha}_{i,x}(t)]}},$$

where $\hat{\alpha}_{i,x}(t) = \frac{1}{K}\sum_{k=1}^{K} g[\mathbf{S}_1^T(t)\hat{\boldsymbol{\nu}}_{x,c} + \mathbf{S}_1^T(t)\hat{\boldsymbol{\Theta}}_{\psi,c}\hat{\boldsymbol{\xi}}_i^{(k)}]$ is an estimate of $P[B_i(t) = 1 \mid \mathbf{B}_i, \mathbf{R}_i]$ and $\hat{\boldsymbol{\xi}}_i^{(k)}$ are Gibbs samples from the final iteration of the MCEM algorithm. The standardized differences for relapse $\varepsilon_{y,c}(t)$ can be similarly defined. We present plots of these statistics in Figure 5. It shows that the standardized differences distribute around 0 with no obvious patterns for both baseline and relapse in all clusters. This indicates that proposed model fits well to the data.

7. DISCUSSION

We consider subtyping cocaine dependent subjects based on their baseline and post-treatment cocaine use behaviors. The data are in the form of paired binary longitudinal cocaine use trajectories before and after a treatment. We propose a novel functional data analysis approach to jointly model and cluster these trajectories. Specifically, we model the non-Gaussian (i.e., binary) response variables by generalized linear mixed models from an exponential family, and model the latent longitudinal processes as functional data. Our approach allows different clusters to have distinct covariance structures, where most existing methods for clustering functional data assume the same covariance structure.

We have identified three clusters based on our cocaine abuse treatment data. The three groups of patients have different probabilities of cocaine use, which suggest differences in the mean structures of the latent processes. Moreover, their covariance structures are also appreciably different. These two observations combined reveal distinct cocaine use behaviors. Specifically, subjects in cluster 1 are heavy users, subjects in cluster 2 use less, but have a strong weekly pattern, and subjects in cluster 3 use cocaine the least often but have both a weekly and decaying correlation structure. Differences in the covariance structures play an important role in defining these clusters, but such differences cannot be detected using any existing methods. We have also found that the relationship between baseline and post-treatment cocaine use behaviors tends to vary across the three clusters.

To balance between model flexibility and complexity, we have made two important dimension reduction efforts in our model. We first register the mean and covariance function by splines, and then further reduce the dimensionality by the much celebrated, data-driven FPCA. We also develop an efficient MCEM algorithm to fit the proposed model. Our simulation and data analysis show that our proposed modeling and estimation procedure can produce insightful results that cannot be obtained by conventional methods.

There is a potential problem of underreporting for self-report drug use data like ours. We relied on the well-established TLFB procedure to construct cocaine use trajectories in our study. To ensure data quality, all staff had been trained by Ph.D.-level psychologists before data collection and were closely supervised when conducting the interviews. All study participants had been informed upfront that all data would be kept strictly confidential and nonidentifiable, and that no legal consequences as protected by law would be incurred to them by providing the most honest answers. The interviews were conducted in a quiet and comfortable testing room, and excellent rapport was established between the staff and the study participants. These measures helped minimize the chance of underreporting, even though this problem could not be completely eliminated.

APPENDIX A: TECHNICAL DETAILS FOR THE E-STEP

We now provide a detailed algorithm for the Gibbs sampler used in the E-step. Given the random samples at the kth step, $(\boldsymbol{\omega}_i^{(k)}, \boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)})$, we take the following steps to update these variables in the $(k+1)$th step.

(A.1). Marginally, $\boldsymbol{\omega}_i \sim \text{Multinomial}[1, (\pi_{1,\text{prev}}, \ldots, \pi_{C,\text{prev}})]$. At the $(k+1)$th step, $\omega_i^{(k+1)}$ is a sample from $[\boldsymbol{\omega}_i|\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}}] \sim \text{Multinomial}[1, (\widetilde{\pi}_{i1}, \ldots, \widetilde{\pi}_{iC})]$ where

$$\widetilde{\pi}_{ic} = P\left(\omega_{ic} = 1|\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}}\right) = \frac{\pi_{c,\text{prev}} A_c(\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}})}{\sum_{c'=1}^{C} \pi_{c',\text{prev}} A_{c'}\left(\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}}\right)}$$

and

$$A_c\left(\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\Omega}_{\text{prev}}\right) = f_c\left(\boldsymbol{\xi}^{(k)}, \boldsymbol{\zeta}^{(k)}|\boldsymbol{\Omega}_{\text{prev}}\right)\left[\prod_{j=1}^{T_i} f_c(B_{ij}|\xi_i^{(k)}, \boldsymbol{\Omega}_{\text{prev}})\right]\left[\prod_{\ell=1}^{S_i} f_c\left(R_{i\ell}|\zeta_i^{(k)}, \boldsymbol{\Omega}_{\text{prev}}\right)\right]$$

(A.2). Next, generate $\boldsymbol{\xi}_i^{(k+1)}$ from

$$f(\boldsymbol{\xi}_i|\boldsymbol{\zeta}_i^{(k)}, \mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}) \propto f(\mathbf{B}_i|\boldsymbol{\xi}_i, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}) f(\boldsymbol{\xi}_i|\boldsymbol{\zeta}_i^{(k)}, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}),$$

which does not have a close form.
Following the Metropolis-Hastings algorithm, we generate a candidate $\boldsymbol{\xi}_i^*$ from $f(\boldsymbol{\xi}_i|\boldsymbol{\zeta}_i^{(k)}, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}})$. Given $\omega_{ic}^{(k+1)} = 1$, this is the density of $\text{Normal}(\boldsymbol{\Lambda}_{c,\text{prev}}\zeta_i^{(k)}, \boldsymbol{\Sigma}_{\eta,c,\text{prev}})$.
Set $\boldsymbol{\xi}_i^{(k+1)} = \boldsymbol{\xi}_i^*$ with probability

$$\alpha_\xi = \min\left\{1, \mathcal{P}_\xi(\boldsymbol{\xi}_i^*)/\mathcal{P}_\xi(\boldsymbol{\xi}_i^{(k)})\right\},$$

where $\mathcal{P}_\xi(\boldsymbol{\xi}_i) = \prod_{j=1}^{T_i} f(B_{ij}|\boldsymbol{\xi}_i, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}})$. If the candidate is rejected, set $\boldsymbol{\xi}_i^{(k+1)} = \boldsymbol{\xi}_i^{(k)}$.

(A.3). Finally, generate $\boldsymbol{\zeta}_i^{(k+1)}$ from

$$f(\boldsymbol{\zeta}_i|\mathbf{B}_i, \mathbf{R}_i, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\xi}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}) \propto f(\mathbf{R}_i|\boldsymbol{\zeta}_i, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}) f(\boldsymbol{\zeta}_i|\boldsymbol{\xi}_i^{(k+1)}, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}).$$

Similar to (10), we can express $\boldsymbol{\zeta}_i$ as

$$\boldsymbol{\zeta}_i = \widetilde{\boldsymbol{\Lambda}}_c \cdot \boldsymbol{\xi}_i + \widetilde{\boldsymbol{\eta}}_{ic},$$

where $\widetilde{\boldsymbol{\Lambda}}_c = \boldsymbol{\Sigma}_{yx,c}\mathbf{D}_{\lambda x,c}^{-1}$, $\widetilde{\boldsymbol{\eta}}_{ic} \sim \text{Normal}(0, \widetilde{\boldsymbol{\Sigma}}_{\eta,c})$, and $\widetilde{\boldsymbol{\Sigma}}_{\eta,c} = \mathbf{D}_{\lambda y,c} - \boldsymbol{\Sigma}_{yx,c}\mathbf{D}_{\lambda x,c}^{-1}\boldsymbol{\Sigma}_{xy,c}$. Generate a candidate $\boldsymbol{\zeta}_i^*$ from $[\boldsymbol{\zeta}_i|\boldsymbol{\xi}_i^{(k+1)}, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}}] \sim \text{Normal}(\widetilde{\boldsymbol{\Lambda}}_{c,\text{prev}}\xi_i^{(k+1)}, \widetilde{\boldsymbol{\Sigma}}_{\eta,c,\text{prev}})$, and set $\boldsymbol{\zeta}_i^{(k+1)} = \boldsymbol{\zeta}_i^*$

with probability

$$\alpha_\zeta = \min\left\{1, \mathcal{P}_\zeta(\boldsymbol{\zeta}_i^*)/\mathcal{P}_\zeta(\boldsymbol{\zeta}_i^{(k)})\right\},$$

where $\mathcal{P}_\zeta(\boldsymbol{\zeta}_i) = \prod_{\ell=1}^{S_i} f(R_{i\ell}|\boldsymbol{\zeta}_i, \boldsymbol{\omega}_i^{(k+1)}, \boldsymbol{\Omega}_{\text{prev}})$. If the candidate is rejected, set $\boldsymbol{\zeta}_i^{(k+1)} = \boldsymbol{\zeta}_i^{(k)}$.

The Gibbs algorithm proceeds by repeating steps (A.1) – (A.3) $\widetilde{K} + K$ times. The first $\widetilde{K}$ samples are discarded as a burn-in period. In the simulation study and data analysis, we use $\widetilde{K} = 500$.

APPENDIX B: TECHNICAL DETAILS FOR THE M-STEP

The detailed algorithm for updating the parameters in the M-step is as follows.

(B.1). Estimation of π.
By maximizing $\widehat{Q}_1$ in (12), we can update the value of π_c by

$$\widehat{\pi}_c = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ic}^{(k)}, \quad c = 1, \ldots, C.$$

(B.2). Estimation of $\boldsymbol{\nu}_{x,c}$, $\boldsymbol{\nu}_{y,c}$, $\boldsymbol{\Theta}_{\psi,c}$ and $\boldsymbol{\Theta}_{\phi,c}$.
We update $\boldsymbol{\nu}_{x,c}$ and $\boldsymbol{\Theta}_{\psi,c}$ by maximizing $\widehat{Q}_2$ with an iteratively reweighted least square (IRLS) method.
We first fix $\boldsymbol{\Theta}_{\psi,c}$, and update $\boldsymbol{\nu}_{x,c}$. Let $\boldsymbol{\nu}_{x,c}^{(h)}$ be the value of $\boldsymbol{\nu}_{x,c}$ at the hth step of IRLS, we can update its value by

$$\boldsymbol{\nu}_{x,c}^{(h+1)} = \boldsymbol{\nu}_{x,c}^{(h)} + [\mathbf{D}(\boldsymbol{\nu}_{x,c}^{(h)})]^{-1}\mathbf{U}(\boldsymbol{\nu}_{x,c}^{(h)}), \qquad (7.15)$$

where $\mathbf{U}(\boldsymbol{\nu}_{x,c}^{(h)}) = \partial\widehat{Q}_2/\partial\boldsymbol{\nu}_{x,c}$ is the gradient of $\widehat{Q}_2$ evaluated at $\boldsymbol{\nu}_{x,c}^{(h)}$ and $\mathbf{D}(\boldsymbol{\nu}_{x,c}^{(h)}) = -\frac{\partial^2\widehat{Q}_2}{\partial\boldsymbol{\nu}_{x,c}\partial\boldsymbol{\nu}_{x,c}^T}$ is the negative Hessian matrix. For the binary case in our data,

$$\mathbf{U}\left(\boldsymbol{\nu}_{x,c}^{(h)}\right) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\omega_{ic}^{(k)}\left[B_{ij} - \alpha_{i,x}^{(k)}(t_{ij})\right]\mathbf{S}_1(t_{ij}),$$

$$\mathbf{D}\left(\boldsymbol{\nu}_{x,c}^{(h)}\right) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\omega_{ic}^{(k)}\alpha_{i,x}^{(k)}(t_{ij})\left[1 - \alpha_{i,x}^{(k)}(t_{ij})\right]\mathbf{S}_1(t_{ij})\mathbf{S}_1^T(t_{ij}),$$

where $\alpha_{i,x}^{(k)}(t) = g[\mathbf{S}^T(t)\boldsymbol{\nu}_{x,c}^{(h)} + \mathbf{S}^T(t)\boldsymbol{\Theta}_{\psi,c}\boldsymbol{\xi}_i^{(k)}]$ and $g(x) = \exp(x)/[1+\exp(x)]$.
Next, put $\boldsymbol{\Theta}_{\psi,c} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{P_{1,c}})$, where $\boldsymbol{\theta}_\ell$ are the spline coefficients for the ℓth eigenfunction. We update the columns of $\boldsymbol{\Theta}_{\psi,c}$ one at a time, while holding all other columns fixed. Let $\boldsymbol{\Theta}_{\psi,c}^{(h)}$ be its value at the hth step, we update the ℓth column by

$$\boldsymbol{\theta}_\ell^{(h+1)} = \boldsymbol{\theta}_\ell^{(h)} + [\mathbf{D}(\boldsymbol{\theta}_\ell^{(h)})]^{-1}\mathbf{U}(\boldsymbol{\theta}_\ell^{(h)}),$$

where $\mathbf{D}(\boldsymbol{\theta}_\ell)$ and $\mathbf{U}(\boldsymbol{\theta}_\ell)$ are the gradient and hessian of $\widehat{Q}_2$ with respect to θ_ℓ. In the binary case,

$$\mathbf{D}\left(\boldsymbol{\theta}_\ell^{(h)}\right) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\omega_{ic}^{(k)}\xi_{i\ell}^{(k)}\left[B_{ij} - \alpha_{i,x}^{(k)}(t_{ij})\right]\mathbf{S}_1(t_{ij}),$$

$$\mathbf{U}(\boldsymbol{\theta}_\ell^{(h)}) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\omega_{ic}^{(k)}\left(\xi_{i\ell}^{(k)}\right)^2\alpha_{i,x}^{(k)}(t_{ij}) \times\left[1 - \alpha_{i,x}^{(k)}(t_{ij})\right]\mathbf{S}_1(t_{ij})\mathbf{S}_1(t_{ij})^T.$$

This procedure is repeated for each column of $\boldsymbol{\Theta}_{\psi,c}$. We iterate it until there is no change in all columns of $\boldsymbol{\Theta}_{\psi,c}$. Denote $\widehat{\boldsymbol{\nu}}_{x,c}$ and $\widetilde{\boldsymbol{\Theta}}_{\psi,c}$ as the results at the convergence of IRLS.
We update $\boldsymbol{\nu}_{y,c}$ and $\boldsymbol{\Theta}_{\phi,c}$ using a similar procedure. Note that $\widetilde{\boldsymbol{\Theta}}_{\psi,c}$ and $\widetilde{\boldsymbol{\Theta}}_{\phi,c}$ are not the estimates of $\boldsymbol{\Theta}_{\psi,c}$ and $\boldsymbol{\Theta}_{\phi,c}$ yet, since the orthonormal constraints in (7) have not been satisfied.

(B.3). Estimation of $\mathbf{D}_{\lambda x,c}$, $\mathbf{D}_{\lambda y,c}$, $\boldsymbol{\Sigma}_{xy,c}$, and orthonormalization of $\widetilde{\boldsymbol{\Theta}}_{\psi,c}$ and $\widetilde{\boldsymbol{\Theta}}_{\phi,c}$.
By maximizing $\widehat{Q}_4$, which is a Gaussian likelihood, we can easily estimate the covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ by

$$\widehat{\boldsymbol{\Sigma}}_{\xi,c} = \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\omega_{ic}^{(k)}\boldsymbol{\xi}_i^{(k)}(\boldsymbol{\xi}_i^{(k)})^T\right]\Big/\left(\sum_{k=1}^{K}\sum_{i=1}^{n}\omega_{ic}^{(k)}\right),$$

$$\widehat{\boldsymbol{\Sigma}}_{\zeta,c} = \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\omega_{ic}^{(k)}\boldsymbol{\zeta}_i^{(k)}(\boldsymbol{\zeta}_i^{(k)})^T\right]\Big/\left(\sum_{k=1}^{K}\sum_{i=1}^{n}\omega_{ic}^{(k)}\right).$$

To estimate $\mathbf{D}_{\lambda x,c}$, $\mathbf{D}_{\lambda y,c}$ and to enforce the orthonormal constraints in (7), define $\boldsymbol{\Gamma}_{\xi,c} = \widetilde{\boldsymbol{\Theta}}_{\psi,c}\widehat{\boldsymbol{\Sigma}}_{\xi,c}\widetilde{\boldsymbol{\Theta}}_{\psi,c}^T$ and $\boldsymbol{\Gamma}_{\zeta,c} = \widetilde{\boldsymbol{\Theta}}_{\phi,c}\widehat{\boldsymbol{\Sigma}}_{\zeta,c}\widetilde{\boldsymbol{\Theta}}_{\phi,c}^T$, where $\widetilde{\boldsymbol{\Theta}}_{\psi,c}$ and $\widetilde{\boldsymbol{\Theta}}_{\phi,c}$ are the estimators in (B.2). We perform an eigenvalue decomposition on $\boldsymbol{\Gamma}_{\xi,c}$ such that $\boldsymbol{\Gamma}_{\xi,c} = \widehat{\boldsymbol{\Theta}}_{\psi,c}\widehat{\mathbf{D}}_{\lambda x,c}\widehat{\boldsymbol{\Theta}}_{\psi,c}^T$, where $\widehat{\mathbf{D}}_{\lambda x,c}$ is a diagonal matrix consisting of the leading $P_{1,c}$ eigenvalues of $\boldsymbol{\Gamma}_{\xi,c}$, and $\widehat{\boldsymbol{\Theta}}_{\psi,c}$ consists of the associated eigenvectors of $\boldsymbol{\Gamma}_{\xi,c}$. $\widehat{\mathbf{D}}_{\lambda x,c}$ and $\widehat{\boldsymbol{\Theta}}_{\psi,c}$ are then the estimates of $\mathbf{D}_{\lambda x,c}$ and $\boldsymbol{\Theta}_{\psi,c}$, respectively. Applying the same procedure to $\boldsymbol{\Gamma}_{\zeta,c}$, we get $\widehat{\mathbf{D}}_{\lambda y,c}$ and $\widehat{\boldsymbol{\Theta}}_{\zeta,c}$.
Finally, define $\widetilde{\boldsymbol{\xi}} = \widehat{\boldsymbol{\Theta}}_{\psi,c}^T\widetilde{\boldsymbol{\Theta}}_{\psi,c}\boldsymbol{\xi}$ and $\widetilde{\boldsymbol{\zeta}} = \widehat{\boldsymbol{\Theta}}_{\phi,c}^T\widetilde{\boldsymbol{\Theta}}_{\phi,c}\boldsymbol{\zeta}$. We can then estimate the cross-covariance matrix between the two sets of FPC scores by

$$\widehat{\boldsymbol{\Sigma}}_{xy,c} = \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\omega_{ic}^{(k)}\widetilde{\boldsymbol{\xi}}_i^{(k)}\left(\widetilde{\boldsymbol{\zeta}}_i^{(k)}\right)^T\right]\Big/\left[\sum_{k=1}^{K}\sum_{i=1}^{n}\omega_{ic}^{(k)}\right].$$

SUPPLEMENTARY MATERIALS

In the supplementary materials, we provide more details of the proposed model selection method, EM algorithm and additional simulation and data analysis results.

[Received March 2013. Revised July 2014.]

REFERENCES

Anton, R. F., O'Malley, S. S., Ciraulo, D. A., Cisler, R. A., Couper, D., and Donovan, D. M. (2006), "Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence The COMBINE Study: A Randomized Controlled Trial," *Journal of the American Medical Association*, 295, 2003–2017. [1417]

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [1413]

Booth, J. G., and Hobert, J. P. (1999), "Maximizing Generalized Linear Mixed Model Likelihoods With an Automated Monte Carlo EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 61, 265–285. [1416]

Carroll, K. C., Power, M., Bryant, K., and Rounsaville, B. J. (1993), "One Year Follow-Up Status of Treatment-Seeking Cocaine Abusers: Psychopathology and Dependence Severity as Predictors Of Outcome," *Journal of Nervous and Mental Disease*, 181, 71–79. [1417]

Crainiceanu, C. M., Staicu, A-M, and Di, C. Z. (2009), "Generalized Multilevel Functional Regression," *Journal of the American Statistical Association*, 104, 1550–1561. [1413]

Chan, K. S., and Ledolter, J. (1995), "Monte Carlo EM Estimation for Time Series Models Involving Counts," *Journal of the American Statistical Association*, 90, 242–252. [1416]

Chiou, J.-M., and Li, P.-L. (2007), "Functional Clustering and Identifying Substructures of Longitudinal Data," *Journal of the Royal Statistical Society*, Series B, 69, 679–699. [1413]

——— (2008), "Correlation-Based Functional Clustering via Subspace Projection," *Journal of The American Statistical Association*, 103, 1684–1692. [1413]

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), "Multilevel Functional Principal Component Analysis," *Annals of Applied Statistics*, 3, 458–488. [1413]

Fals-Stewart, W., O'Farrell, T. J., Freitas, T. T., McFarlin, S. K., and Rutigliano, P. (2000), "The Timeline Follow-Back Reports of Psychoactive Substance Use by Drug-Abusing Patients: Psychometric Properties," *Journal of Consulting and Clinical Psychology*, 68, 134–144. [1412]

Fox, H. C., Garcia, M., Milivojevic, V., Kreek, M. J., and Sinha, R. (2006), "Gender Differences in Cardiovascular and Corticoadrenal Response to Stress and Drug Cues in Cocaine Dependent Individuals," *Psychopharmacology*, 185, 348–357. [1419]

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, 6, 721–741. [1416]

Guan, Y., Li, Y., and Sinha, R. (2011), "Cocaine Dependence Treatment Data: Methods for Measurement Error Problems With Predictors Derived From Stationary Stochastic Processes," *Journal of the American Statistical Association*, 106, 480–493. [1413,1414]

James, G., Hastie, T., and Sugar, C. (2000), "Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587–602. [1413,1414]

James, G., and Sugar, C. (2003), "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association*, 98, 397–408. [1413]

Jiang, H., and Serban, N. (2011), "Clustering Random Curves Under Spatial Interdependence With Application to Service Accessibility," *Technometrics*, 54, 108–119. [1413]

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. [1413]

Hall, P., Müller, H. G., and Yao, F. (2008), "Modelling Sparse Generalized Longitudinal Observations With Latent Gaussian Processes," *Journal of the Royal Statistical Society*, Series B, 70, 703–723. [1413,1415]

Hubert, L. J., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [1417]

Ibrahim, J. G., Zhu, H., and Tang, N. (2008), "Model Selection Criteria for Missing-Data Problems Using the EM Algorithm," *Journal of the American Statistical Association*, 103, 1648–1658. [1416]

Li, T. (2006), "A Unified View on Clustering Binary Data," *Machine Learning*, 62, 199–215. [1413]

Li, Y., and Hsing, T. (2010), "Uniform Convergence Rates for Nonparametric Regression and Principal Component Analysis in Functional/Longitudinal Data," *The Annals of Statistics*, 38, 3321–3351. [1413,1417]

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170. [1416]

McCulloch, P., and Nelder, J. (1989), *Generalized Linear Model* (2nd ed.), London: Chapman & Hall. [1414]

Ordonez, C. (2003), "Clustering Binary Data Streams With k-Means," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, eds. M.J. Zaki and C.C. Aggarwal, DMKD 2003, San Diego, CA,12–19. [1413]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer-Verlag. [1413]

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850. [1417]

Rao, J. N. K., and Scott, A. J. (1992), "A Simple Method for the Analysis of Clustered Binary Data," *Biometrics*, 48, 577–585. [1413]

Serban, N., and Jiang, H. (2012), "Multilevel Functional Clustering Analysis," *Biometrics*, 68, 805–814. [1413]

Sinha, R., Garcia, M., Paliwal, P., Kreek, M. J., and Rounsaville, B. J. (2006), "Stress-Induced Cocaine Craving and Hypothalamic-Pituitary-Adrenal Responses are Predictive of Cocaine Relapse Outcomes," *Archives of General Psychiatry*, 633, 324–331. [1417]

Sobell, L., and Sobell, M. (1993), "Timeline Follow Back: A Technique for Assessing Self-Reported Ethanol Consumption," in *Techniques to Assess Alcohol Consumption*, eds. J. Allen and R. Litten, Totowa, NJ: Humana Press, Inc, 41–42. [1412]

Tanner, M. A. (1993), *Tools for Statistical Inference: Observed Data and Data Augmentation* (2nd ed.), Berlin: Springer-Verlag. [1416]

Wei, C. G., 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. [1416]

Yao, F., Müller, H. G., and Wang, J. L. (2005a), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [1413]

——— (2005b), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903. [1413,1415]

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008), "Joint Modelling of Paired Sparse Functional Data Using Principal Components," *Biometrika*, 95, 3, 601–619. [1413,1414,1415]

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. [1413]