Published in final edited form as:

Stat Biosci. 2019 July ; 11(2): 238–261. doi:10.1007/s12561-019-09232-8.

Clustering functional data with application to electronic medication adherence monitoring in HIV prevention trials

Yifan Zhu,

1100 Fairview Ave. N., Seattle WA, USA, Tel.: +1-206-6674526, yzhu2@fredhutch.org

Chongzhi Di,

1100 Fairview Ave. N., Seattle WA, USA, Tel.: +1-206-6677004, cdi@fredhutch.org

Ying Qing Chen

1100 Fairview Ave. N., Seattle WA, USA, Tel.: +1-206-6677051, yqchen@fredhutch.org

Abstract

Maintaining high medication adherence is essential for achieving desired efficacy in clinical trials, especially prevention trials. However, adherence is traditionally measured by self-reports that are subject to reporting biases and measurement error. Recently, electronic medication dispenser devices have been adopted in several HIV pre-exposure prophylaxis prevention studies. These devices are capable of collecting objective, frequent, and timely drug adherence data. The device opening signals generated by such devices are often represented as regularly or irregularly spaced discrete functional data, which are challenging for statistical analysis. In this paper we focus on clustering the adherence monitoring data from such devices. We first pre-process the raw discrete functional data into smoothed functional data. Parametric mixture models with change-points, as well as several non-parametric and semi-parametric functional clustering approaches are adapted and applied to the smoothed adherence data. Simulation studies were conducted to evaluate finite sample performances, on the choices of tuning parameters in the pre-processing step as well as the relative performance of different clustering algorithms. We applied these methods to the HIV Prevention Trials Network(HPTN) 069 study for identifying subgroups with distinct adherence behavior over the study period.

Keywords

Drug adherence; HIV prevention; Clustering; Functional data; Latent class model

1 Introduction

1.1 Motivating study: HPTN 069 Study

Adherence plays an essential role in evaluating efficacy in clinical trials, especially in prevention trials. However, it is challenging to measure adherence accurately. Traditionally,

adherence levels were measured by self-reported questionnaires (e.g., on pill taking frequency). However, these data are subject to reporting biases and low resolution (i.e., only a few times during the study period). In recent years, advances in technology have made it possible to measure adherence objectively by electronic medication monitoring devices. For example, Wisepill is an electronic medication dispenser device, which sends a signal to the data collection center via cellular network when the device is opened [9]. Data recorded by these devices not only provide objective measurements of adherence, but also allow assessment of temporal changes of adherence over time.

This study was directly motivated by the HIV Prevention Trials Network(HPTN) 069 study [8], a phase II randomized, double-blind study of safety and tolerability of four active-arm pre-exposure prophylaxis(PrEP) for HIV infection prevention. The active arm daily prescriptions were maraviroc (MVC), maraviroc + emtricitabine (MVC+FTC), maraviroc + tenofovir disoproxil fumarate (MVC+TDF) or tenofovir disoproxil fumarate + emtricitabine (TDF+FTC) for high-risk HIV populations including men who have sex with men (MSMs) and at-risk women. The study enrolled a total of 594 subjects from 13 sites in the United States and Puerto Rico, consisting of 406 MSMs and 188 females who had been followed-up for 350 days after randomization. Study subjects are expected to take medications daily in the HPTN 069 trial. Wisepill dispensers were used to track device opening continuously. Data collected for each individual includes timestamps of device opening signals, over the study period of 350 days. The devices also sent daily heartbeat signals to indicate that they were working properly. Missing Wisepill records were present mainly due to device failures such as depleted battery, incompatible cellular network for the SIM cards inserted into Wisepill device, or other unknown reasons.

The subject drug adherence levels in HPTN 069 were also measured by computer assisted self interviewing (CASI) questionnaires. The CASI was performed during study follow-up visits every 8 weeks, asking subjects about their self-reported frequency of pill taking behaviour in the past 4 weeks. These type of measurements were more subjective and often overestimate the actual adherence levels ([7]). Being a relatively low-cost, unobtrusive, more objective and frequent adherence monitoring instrument, Wisepill can objectively record individual drug taking patterns in high temporal resolution.

1.2 Adherence as functional data

The main objectives for statistical analysis are to understand and cluster patterns of adherence and to identify personal characteristics that are associated with various patterns. Wisepill data can be viewed as discrete functional data, as data for each subject is a function that reflects daily adherence levels. The major challenges for clustering Wisepill adherence include high observation frequency, irregular time points, missing records and most importantly, the binary data type. Due to these properties of Wisepill adherence data, routine clustering approaches including multivariate methods, mixture distribution models and functional clustering algorithms require special attention with the current application. We propose to adapt existing methods for clustering functional data, including both parametric and non-parametric methods, and to apply them to the HPTN 069 Study. Additionally, we propose a semi-parametric approach which incorporates parametric components for

improved interpretability and flexibly models covariance structure nonparametrically, which could be advantageous for this application. To our best knowledge, this paper is among the first attempts on clustering generalized functional data with a novel application in mobile health settings.

The rest of the paper is organized as follows. Section 2 introduces notations and pre-processing procedures to transform data into continuous functional data. Section 3 discusses several parametric, non-parametric as well as semi-parametric methods for functional clustering. Simulation studies to evaluate the finite sample performance of these methods are presented in Section 4, while Section 5 illustrates their application to the HTPN 069 study.

2 Data, notations and pre-processing

The raw Wisepill data includes timestamps of device opening signals. The data were transformed into a binary time-series over 350 days for each individual, denoted as $\mathbf{W}_i = \{W_{il}: l = 1, 2, \dots, L_i\}$ recorded at time points $\{t_{i1}, \dots, t_{iL_i}\}$ for subject $i, i \in \{1, 2, \dots, n\}$. $W_{il} = 0$ means the subject did not open the pill box while $W_{il} = 1$ means the subject opened the pill box at least once on day t_{il} .

Let $A_i(t)$ denote the continuous underlying adherence process for the i^{th} subject. The aim of the analysis is to cluster patterns of adherence processes. However, since it is unobservable, we first pre-smooth the observed binary process $\mathbf{W}_i$ by kernel smoothing or local moving average smoothing. For example, at the current time t , one can consider a symmetric time window $(t-d, \dots, t, \dots, t+d)$ with length $2d+1$, or a historical time window $(t-2d, t-2d+1, \dots, t)$ with the same length. The smoothed adherence process $Y_i(t)$ at time t is defined by $Y_i(t) = \frac{1}{2d+1} \sum_{l=0}^{2d} a_l W_{i(t-l)}$ for historical time windows, or $Y_i(t) = \frac{1}{2d+1} \sum_{l=-d}^d a_l W_{i(t+l)}$ for symmetric time windows, where a_l 's are pre-specified weights determined by the type of kernels for smoothing.

For example, choosing symmetric time windows and constant weights implies that $Y_i(t)$ can be interpreted as the average pill device opening percentage around the t^{th} day. On the other hand, if one chooses to use historical time windows and weights derived from Pharmacokinetic/pharmacodynamic (PK/PD) knowledge, often with exponentially decayed

weights such as $a_l = \frac{e^{-\lambda l}}{\sum_k e^{-\lambda k}}$ (see eg. [6]), $Y_i(t)$ can be interpreted as estimated total drug

concentration levels on the t^{th} day resulting from certain pill taking trajectories. However, it is worth noting that these choices correspond to the very different interpretations of the time location parameter t .

After pre-processing, the smoothed adherence process $Y_i(t)$ will be used in the subsequent analysis. Figure 1 displays example daily Wisepill signal patterns and their corresponding smoothed adherence level patterns. The left panel is a lasagna plot where each row illustrates one study subject's binary daily Wisepill signals. For a specific follow up day l , orange color means $W_{il} = 1$ and pink color means $W_{il} = 0$, while white means the signal is missing. The right panel shows an example of smoothed adherence patterns for each subject from the left

panel, where symmetrical smoothing windows with length equal to 57 days were applied for moving average smoothing and constant weights were used.

3 Methods for clustering functional data

We consider both parametric and non-parametric clustering methods for functional data. For example, when prior knowledge on the adherence process is available, parametric models that take advantage of this knowledge often provide better goodness of fit and interpretability. In the context of medication adherence, Vrijens [24] combined phase IV clinical studies for once a day anti-hypertensive medication for 1989–2006, and found that with electronically compiled dosing history by a medication event monitor, “after the sharp small initial drop of non-acceptance, the persistence curve decreases gradually but progressively over time.” The aggregated and degree of execution stratified adherence time course generally illustrated two stages of approximately linear decreasing pattern with the change of slope occurring at about 50 days after enrollment over one year follow-up ([5]). Therefore, we first propose change-point based mixture models for the average drug taking adherence pattern over time in the current study.

While easy to interpret, parametric mixture methods have drawbacks of relying on the assumption that true adherence process follows the pre-specified parametric distribution families. Non-parametric functional clustering methods, on the other hand, are more flexible with any type of underlying process shape. In the following, we adapt both parametric and non-parametric methods to cluster Wisepill adherence data.

3.1 Parametric change-point mixture models

We first propose a parametric mixture model, based on piecewise linear trend with a change point. We assume that Y_{ij} 's are generated from a mixture model with C total clusters. For each subject i , define cluster membership $\eta_i \in \{1, 2, \dots, C\}$. Assume that η_i follows multinomial distribution, $P(\eta_i = c) = \pi_c$ where π_c is the cluster mixing probability satisfying $\sum_{k=1}^C \pi_k = 1$.

If subject i belongs to cluster c , the proposed model assumes that its adherence process follows a change-point model:

$$E\{Y_i(t) \mid \eta_i = c; \mu_0^c, t_0^c, \beta_1^c, \beta_2^c\} = \mu_0^c + \beta_1^c t + \beta_2^c t I(t \geq t_0^c). \quad (1)$$

The model comprises two linear pieces connected at time t_1^c with changing slopes from β_1^c to $\beta_1^c + \beta_2^c$. Therefore, when latent mixing component of the smoothed Wisepill record is

denoted as $Y_i^c(t) = \mu_0^c + \beta_1^c t + \beta_2^c t I(t \geq t_0^c) + \varepsilon_c(t)$ with random Gaussian noise term

$\varepsilon_c(t) \sim N(0, \sigma_c^2(t))$. For simplicity we assume time-constant measurement error: $\sigma_c(t) = \sigma_c$.

Since the Wisepill smoothing process is performed at daily basis, the subject i 's l^{th} available smoothed data point is $Y_{il} = Y_i^c(t_{il})$, where $l = 1, \dots, L_i$. Thus the likelihood for observing the whole process $\mathbf{Y}_i$ is:

$$f(\mathbf{Y}_i | \boldsymbol{\theta}) = \sum_{c=1}^C \pi_c \prod_{l=1}^{L_i} \Phi(Y_{il} | \eta_i = c; \theta_c, \sigma_c^2),$$

where $\theta_c = (\mu_0^c, t_0^c, \beta_1^c, \beta_2^c)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_C)$, and $\Phi(\cdot)$ is the density function of standard normal distribution.

To maximize the likelihood function $L(\boldsymbol{\theta}) = \prod_i f(\mathbf{Y}_i | \boldsymbol{\theta})$ in the presence of latent class membership η_i , it is natural to employ the expectation-maximization (EM) algorithm. The expectation step involves calculating the conditional expectation of the complete data log-likelihood conditional on observed data $\mathbf{Y}_i$, given the current parameter value $\boldsymbol{\theta}^0$,

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0) = \sum_i E_{\boldsymbol{\theta}^0} [\log f(\mathbf{Y}_i, \eta_i; \boldsymbol{\theta}) | \mathbf{Y}_i],$$

which can be calculated once we obtain the conditional probabilities

$$P(\eta_i = c | \mathbf{Y}_i; \boldsymbol{\theta}^0) = \frac{\pi_c \prod_l \Phi(Y_{il} | \theta_c^0)}{\sum_j \pi_j \prod_l \Phi(Y_{il} | \theta_j^0)}.$$

The next step is to maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0)$ with respect to $\boldsymbol{\theta}$. One can show that the solution satisfy

$$\begin{aligned} \pi_k &= \frac{\sum_i P(\eta_i = c | \mathbf{Y}_i; \boldsymbol{\theta}^0)}{n} \\ \sum_i P(\eta_i = c | \mathbf{Y}_i; \boldsymbol{\theta}^0) \frac{\partial}{\partial \theta^k} \sum_l \log \Phi(Y_{il} | \theta) &= 0, \end{aligned}$$

which can be solved by optimizer algorithms.

The overall EM algorithm for estimating parameters can be summarized as follows:

1. Initialize parameters $\boldsymbol{\theta}^0, \boldsymbol{\pi}^0$.
2. During the k^{th} iteration,
 - (a). E-Step: calculate the complete data likelihood $L(\mathbf{Y} | \boldsymbol{\theta}^{ite})$ using the current parameter values $\boldsymbol{\theta}^{k-1}$.
 - (b). M-Step: update parameter estimates, including the mixing probabilities and β parameters, by using BFGS(a quasi-Newton method with good performance with non-smooth optimization problems [17]) or simulated annealing (uses no derivative from function being optimized [2]) numerical maximization algorithms to obtain updated model parameters $\boldsymbol{\theta}^{k+1}$.

- (c). Evaluate convergence of the algorithm by checking whether changes in the log-likelihood and parameter estimates are smaller than a pre-specified threshold, i.e., $|L^{k+1} - L^k| < \delta$, $|\pi^{k+1} - \pi^k| < \delta$, etc. If the criterion is satisfied, go to Step 3. Otherwise, set $k = k + 1$ and return to step 2.

3. Return the parameter estimates of the final iteration.

As a byproduct, the algorithm also provides estimated posterior class memberships, $P(\eta_i = c | Y_i; \theta_{final})$, which can then be used to assign each subject into a specific class by the largest posterior probability. Standard errors of the parameter estimates can be obtained by calculating the Fisher information matrix [18].

The maximized likelihood evaluated at the EM estimates can be used to construct model-selection criteria such as AIC or BIC. These criteria may be used to evaluate the goodness-of-fit of the current model, such as to determine the total number of clusters C . Although the model-selection procedure is actually dealing with non-nested models with different choices of C , AIC or BIC remain useful candidates.

3.2 Functional clustering

An alternative non-parametric approach of analyzing densely recorded, high-dimensional time series data is through functional data analysis. Assume $Y_{ij} = X_j(t) + \varepsilon_{ij}$, where Y_{ij} 's are observed as dense realizations of a random function $X(t)$ with random i.i.d noise ε_{ij} with mean 0 and variance σ^2 . Non-parametric functional methods do not make assumptions on the shapes of the true process $X(t)$ except that it is smooth. We will briefly review a few popular functional clustering methods in this section.

For a set of pre-specified basis functions $\{\phi_k(t) | k = 1, 2, \dots\}$, $X(t)$ can be represented by

$$X_i(t) = \mu(t) + \sum_k \xi_{ik} \phi_k(t),$$

where $\mu(t)$ is the overall mean function, and ξ_{ik} 's are random variables with zero means.

Commonly used basis include B-splines, wavelets or orthogonal eigenfunctions of the covariance kernel $C(\cdot, \cdot)$ of $X(t)$. If one chooses to use principal component bases, the coefficients ξ_{ik} 's are uncorrelated with zero mean and variance λ_k 's, where $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ are the eigenvalues of $C(\cdot, \cdot)$. In this case, ξ_{ik} 's are called functional principal component scores and $\phi_k(t)$'s are called eigenfunctions.

By restricting the basis dimension to a finite positive integer q , the approximation of functional expansion is:

$$X_i^{(q)}(t) = \mu(t) + \sum_{k=1}^q \xi_{ik} \phi_k(t).$$

Thus the resulting reduced-dimension random coefficient vector $(\xi_{i1}, \dots, \xi_{iq})$ for the i th curve could be fed into standard k-means algorithm for clustering. This approach assumes different clusters share the same projection space generated from a common function basis with same dimension q . One approach for clustering is a two-stage method, based on dimension reduction first (e.g., through functional principal component analysis) and then apply standard clustering methods for multivariate data (coefficient vector). This method is relatively simple, however, it has been shown that its performance is often inferior to the following one-step methods.

3.2.1 Brief review of functional clustering—Traditional clustering algorithms were initially developed in finite dimensional settings, where observations are often multivariate random vectors. The most commonly used algorithms, such as hierarchical clustering [25] or k-means [11], used heuristic and geometric procedures. Probability density summaries of random vectors were used in popular clustering methods such as model-based Gaussian clustering [4]. In the case of functional data that are infinite dimensional, several difficulties hinder the direct usage of methods developed for multivariate data. These difficulties include, for example, the lack of a proper definition of probability density of functional data or distances between observations with noisy inputs.

Jacques and Preda [14] provided an overview of a few distinct types of functional clustering methods for continuous functional data. The first is a two-stage approach, which performs dimension reduction commonly by projection into a set of basis functions in the first stage, and clustering the expansion coefficient vectors in the second stage. There are several choices of different functional basis for expansion (e.g., B-splines or functional principal component basis), as well as the degree of approximation of curve mean and variation processes. The second approach often consists of non-parametric approaches that rely on definition of distances/dissimilarities between functional observations and apply geometric measure based clustering with the obtained measures. To reduce the impact of noisy observations/irregular data, sometimes expected distances instead of the raw distances are used. The third approach considers probability models for the approximation vectors obtained in the “two-stage” methods, thus performing dimension reduction and clustering at the same time.

In order to perform clustering on Wisepill data that are discrete and high-dimensional, commonly used algorithms such as k-means or integer-vector clustering methods are not suitable. Furthermore, functional clustering methods mentioned above were developed for continuous functional observations, and thus are not directly applicable to discrete functional data. Additionally, only a few previous studies have compared the clustering performance among different functional clustering approaches [14].

3.2.2 Model-based clustering—James and Sugar [15] proposed a one-step approach by assuming that the basis coefficients are generated from mixture Gaussian distributions with cluster specific parameters, instead of feeding these coefficients into vector clustering algorithms. They choose to expand each individual functional into spline spanned spaces $\sigma\{s_k(t) | k = 1, \dots, K\}$ with potentially irregular time grid $t_{i1}, \dots, t_{iM_i}$.

$$Y_{i,l} = \sum_{k=1}^q s_k(t_{il})\eta_{ik} + \varepsilon_{i,l}.$$

The basis coefficients $\eta_i = (\eta_{i1}, \dots, \eta_{iq})$ can be expressed by $\eta_i = \mu_{z_i} + \gamma_i$, where μ_{z_i} is cluster specific coefficient mean and γ_i are the deviation from mean for i th curve. Y_i is identified as belonging to z_i th cluster, $z_i \in (1, \dots, G)$. By further assuming cluster specific subspace dimension reduction with linear projection operator $\Lambda_{q \times h}$, $h = \min(q, G-1)$ such that $\mu_k = \lambda_0 + \Lambda \alpha_k$, $\varepsilon \sim N(0, \sigma)$ and $\gamma_i \sim N(0, \Gamma)$, with $M_i \times q$ -dimensional matrix S_i being spline basis evaluated on grid $t_1, \dots, t_{M_i}$:

$$\mathbf{Y}_i | z_i = k \sim N(S_i(\lambda_0 + \Lambda \alpha_k), \Sigma_i = \sigma^2 I + S_i \Gamma S_i^T).$$

The complete data likelihood with model parameter $\Theta = (\Gamma, \sigma, \lambda_0, \Lambda, \alpha)$ and cluster mixing probability $\pi_k, k = 1, \dots, G$ is:

$$L(\Theta, \pi | \mathbf{Y}_1, \dots, \mathbf{Y}_n) = \prod_{i=1}^n \sum_{k=1}^G \pi_k f_k(\mathbf{Y}_i | \theta_k),$$

which is maximized by the E-M algorithm. Clustering is thus performed with posterior membership probabilities $p(z_i = k | \mathbf{Y}_i, \hat{\Theta})$, $k = 1, \dots, G$. We refer this approach as Fclust functional clustering.

3.2.3 Distance-based functional clustering—There are also a few distance-based clustering approaches for functional data. For example, Peng and Muller [19] proposed an approach that defines the distance between two functional observations $\mathbf{Y}_i$ and $\mathbf{Y}_j$ as the L^2 -norm of the difference between their generating processes: $D(\mathbf{Y}_i, \mathbf{Y}_j) = \{\int_T (X_i(t) - X_j(t))^2 dt\}^{1/2}$. The expected distance is defined when the observations are based on sparse grid however:

$$\tilde{D}(Y_i, Y_j) = \{E[D^2(\mathbf{Y}_i, \mathbf{Y}_j) | \mathbf{Y}_i, \mathbf{Y}_j]\}^{1/2} \quad \text{for } 1 \leq i, j \leq n.$$

If one adopts functional principal component decompositions of the processes, the distance defined above can also be approximated by

$$E[D^2(\mathbf{Y}_i, \mathbf{Y}_j) | \mathbf{Y}_i, \mathbf{Y}_j] = E\left[\sum_{k=1}^q (\xi_{ik} - \xi_{jk})^2 | \mathbf{Y}_i, \mathbf{Y}_j\right], \quad (2)$$

as long as the first q principal components explain most of the variation for the true process. Peng and Muller [19] evaluated equation (2) explicitly by multidimensional scaling to configure the curves in a 2-dimensional space. The evaluated distances thus can be used as input to perform k-means clustering. We refer this approach as distFPCA functional clustering.

Another clustering approach, k-centers functional clustering, from Chiou and Li [3] is based on iteratively assigning an observed curve to the closest subspace, by the distance measure between the corresponding projection of the curve into the subspace and the curve itself. The clusters are defined by these subspaces, which can be spanned by different sets of basis functions. Suppose the observations have been initially clustered by a simple algorithm such as k-means into C clusters. For the cluster $c \in 1, \dots, C$, it has a specific cluster mean $\mu^c(t)$, and basis functions $\phi_k^c(t)$. Then in each iteration, the k-centers assign the curve $Y_\lambda(t)$ into new cluster in the following steps:

- With leave-one-out step, re-estimate the subspace mean function $\mu_{-i}^c(t)$ and basis functions $\phi_{j,-i}^c(t)$'s for the cluster c^0 which $Y_\lambda(t)$ previously belongs to, using the remaining observations in that cluster. For other clusters their mean and basis functions remain the same.
- Perform Karhunen-Loeve (K-L) expansion of the curve at each cluster, to estimate the projection of $Y_\lambda(t)$ into all subspaces. For $c = c^0$,

$$\hat{Y}_{-i}^c(t) = \mu_{-i}^c(t) + \sum_{k=1}^{q_c} \xi_{ik}^c \phi_{k,-i}^c(t),$$

and for other clusters,

$$\hat{Y}^c(t) = \mu^c(t) + \sum_{k=1}^{q_c} \xi_{ik}^c \phi_k^c(t).$$

- Compute the L^2 distance between the observation and its current subspace projections, and assign $Y_\lambda(t)$ to cluster c^* such that

$$c^*(y) = \operatorname{argmin}_c \|Y - \hat{Y}^c\|_{L^2}.$$

Repeat these steps for all i iteratively until no new assignment is needed. The k-centers clustering was originally developed for regularly observed data. However the conditional expectation approach (e.g., see [26]) for the distance estimation in distFPCA could be adapted here to adapt k-centers functional clustering (FC) for irregularly spaced data. Several parameters can be tuned in the procedure for application, including the type of basis functions, basis dimension q_c in each cluster, and whether to perform the re-estimation of the mean and basis functions in cluster c^0 .

3.3 Semi-parametric change-point mixture models

In the change-point mixture models described in section 3.1, we have assumed that $\varepsilon_c(t) \sim N(0, \sigma_c^2)$ for $t = t_1, \dots, t_L$. This implies that the with-in subject measurements at different time points are independent and they share common variance. Such simplified assumption might be inadequate for complex covariance structures. In this section we present a more

general model with increased flexibility for correlated measurements over time. Let $\Sigma_c(s, t) = \text{Cov}(\mathbf{e}_{cs}, \mathbf{e}_{ct})$ be the covariance process of individual $Y_c(t)$'s in cluster c . We can imply the flexible structure for Σ_c by using functional principal component estimation with

$$\hat{\Sigma}_c(s, t) = \sigma_c^2 I + \sum_{k=1}^{p_c} \lambda_k^c \Phi_k^c(s) \Phi_k^c(t).$$

Here σ_c^2 accounts for the measurement error variance, and $\Phi_k^c(t)$ is the k^{th} eigenfunction for the residual covariance process with corresponding eigenvalue λ_k^c , all defined within cluster c . These parameters and basis functions can be partially shared among clusters to allow flexibility in modeling heterogeneity levels. We truncate the K-L expansion of residual covariance surface to p_c eigenfunction basis dimension, with p_c often chosen by preserving 90% of total variation. The cluster-specific covariance structure estimated this way thus provides more flexible non-parametric representation than previously used $\sigma_c^2 I$. The likelihood of observing $\mathbf{Y}_i$ thus becomes:

$$f(\mathbf{Y}_i | \theta, \Sigma) = \sum_{c=1}^C \pi_c \Phi_{MVN}(Y_{i1}, \dots, Y_{iL_i} | \eta_i = c; \theta_c, \hat{\Sigma}_c),$$

where $\Phi_{MVN}(\cdot)$ is the multivariate normal density function. The mean vector and the covariance matrix of the multivariate normal density are defined by the change-point model with parameter θ_c and $\hat{\Sigma}_c$ evaluated at observation time grids $(t_{i1}, \dots, t_{iL_i})$. The semi-parametric clustering model is an extension of the change-point mixture model that combines the parametric model-based mean process with non-parametric unstructured covariance process for the distribution of the cluster members.

We sketch our semi-parametric clustering procedure in the following steps:

1. Perform change-point EM clustering, obtaining subject cluster membership η_i^{EM} 's. Set $\theta^0 = \theta^{EM}$ and $\pi^0 = \pi^{EM}$. Estimate Σ_c^0 from the residual processes $r_i(t) = Y_i(t) - f_{c=\eta_i^{EM}}(t | \theta^c = \eta_i^{EM})$ with functional principal component analysis (FPCA), with the data irregularity handled by local linear smoothing described in [10].
2. During the k^{th} iteration, perform model optimization, iteratively
 - update θ^k and π^k by maximizing the profile likelihood when fixing $\Sigma = \Sigma^{k-1}$, similar to the EM algorithm described previously in section 3.1.
 - estimate $\sigma_c^{k-0.5}$'s with $r_i^{k-0.5}(t) = Y_i(t) - f_{c=\eta_i^{k-1}}(t | \theta^k)$ by FPCA.

- re-estimate $\eta_i^{k-0.5}$'s with posterior probabilities

$$P(\eta_i = c \mid \mathbf{Y}_i; \boldsymbol{\theta}^k, \Sigma^{k-0.5}) = \frac{\pi_c^k \Phi_{MVN}(Y_{i1}, \dots, Y_{iL_i} \mid \boldsymbol{\theta}_c^k, \Sigma_c^{k-0.5})}{\sum_j \pi_j^k \Phi_{MVN}(Y_{i1}, \dots, Y_{iL_i} \mid \boldsymbol{\theta}_j^k, \Sigma_j^{k-0.5})}.$$
- update Σ_c^k 's from $r_i^k(t) = Y_i(t) - f_{c = \eta_i^{k-0.5}(t \mid \boldsymbol{\theta}^k)}$ by FPCA.
- update η_i^k 's with posterior probabilities

$$P(\eta_i = c \mid \mathbf{Y}_i; \boldsymbol{\theta}^k, \Sigma^k) = \frac{\pi_c^k \Phi_{MVN}(Y_{i1}, \dots, Y_{iL_i} \mid \boldsymbol{\theta}_c^k, \Sigma_c^k)}{\sum_j \pi_j^k \Phi_{MVN}(Y_{i1}, \dots, Y_{iL_i} \mid \boldsymbol{\theta}_j^k, \Sigma_j^k)}.$$
- Convergence is checked using two criteria: (i). the stability of clustering measured by the difference between $\boldsymbol{\eta}^k$ and $\boldsymbol{\eta}^{k-1}$, where a lower threshold of Rand Index([20]) between them is set to be $RI_{hybrid} = 0.99$; (ii). The L^2 -norm type difference between $(\boldsymbol{\theta}^k, \boldsymbol{\pi}^k, \Sigma^k)$ and $(\boldsymbol{\theta}^{k-1}, \boldsymbol{\pi}^{k-1}, \Sigma^{k-1})$, with upper limit δ_{hybrid} . If both criteria are met, go to Step 3. Otherwise, set $k = k + 1$ and return to step 2.

3. Return the parameter and cluster membership estimates from the final iteration.

The final clustering is determined using the posterior cluster membership probabilities evaluated at converged model parameters. Similar to the change-point models, individuals are assigned to the cluster with largest posterior probability.

3.4 How many clusters?

The topic of choosing the number of clusters has been discussed in many of the functional clustering literature mentioned in this section. Most clustering algorithms require fixing the total cluster sizes before fitting, and it is one of the most difficult problems in cluster analysis ([15]). One may determine the number of clusters heuristically by fitting the data to several possible cluster sizes and seeing whether the results produce physical meanings ([3]). Several formal and systematic approaches for total cluster number determination have been discussed. For model-based clustering, model-selection criteria such as Bayes factors ([16]) or its approximation BIC ([21]) as well as AIC ([1]) and their adaption to specific clustering models were frequently used. Quantities to measure similarity within clusters and dissimilarity across different clusters have been developed to evaluate the clustering result, for example the “distortion function” ([15]). Also, measure for prediction stability from cross-validation approach has been proposed by Tibshirani and Walther [22].

4 Simulation studies

We conducted simulation studies to evaluate the performance of various clustering methods in dense and discrete functional data that mimic the Wisepill adherence data from our motivating study. Although finite sample performance for some of the clustering methods have been assessed for continuous functional data, it is unclear how they would perform as part of two-step procedures in clustering discrete functional data.

4.1 Clustering Performance Comparison of Parametric vs. Functional Approaches

We start from simulating the latent true drug adherence time series $A_i(t)$ for subjects $i = 1, \dots, n$. The drug adherence is followed up during $t \in [1, 150]$ continuously. In each repeated simulation, we assume there are three clusters with distinct mean process $f_{sim}(t)$ and random error process $\varepsilon_{sim}(t)$. We consider two scenarios for change-point assumption. For the scenario 1 and 2 where the change-point mean assumption holds, each cluster specific mean process follows one of five possible candidate parameter vectors $\theta = (\mu, \beta_1, \beta_2, t_0)$ according to equation 1: $f_{sim,i}(t)|c_i = f_{c,i}(t|\theta)$. Let $\mu_0 = 1$ and $\beta_2 = 0$ for all five candidates such that all mean curves start from full adherence and remain flat after change-point t_0 ; the remaining parameter scenarios are:

$$(\beta_1, t_0) \in \{(0.02, 30), (0.002, 200), (0.005, 120), (0.03, 10), (0.001, 100)\}.$$

For scenario 3 where the change-point assumption is violated, we add two additional candidates to the previous five change-point model candidates for the mean process: $f_1^*(t) = 0.7 + 0.1 \sin(\pi t / 60)$ and $f_2^*(t) = 0.7 + 0.1 \sin(\pi t / 30 - 0.002t)$. The simulated true latent adherence $A_i(t) = f_{sim,i}(t)|c_i + \varepsilon_{sim,i}(t)$, where $\varepsilon_{sim,i}(t) \sim N(0, \sigma_{c_i}^2)$ and $\sigma_{c_i} \sim N(0.1, 0.01^2)$.

Following simulated $A_i(t)$, subject i 's complete daily Wisepill records are then generated with $W_i^0(t) \sim \text{Bin}(1, A_i(t))$ for $t = 1, \dots, 150$. Denote $M_i(t)$ as the missing indicator, we assume $M_i(t) \sim \text{Bin}(1, p_m)$ where p_m is the unconditional missing probability. The observed Wisepill record $W_i(t) = W_i^0(t)$ if $M_i(t) = 0$ and $W_i(t)$ is missing if $M_i(t) = 1$. p_m is set to equal to 0.3 for all $i = 1, \dots, n$ and $t = 1, \dots, 150$.

The pre-smoothing for simulated $W_i(t)$ is performed with exponentially weighted moving window averaging, where the window length d is chosen to be $d \in \{7, 14, 28, 56\}$. The type of moving window could be historical window where $d = L$, or symmetrical window where $d = C$. The exponential weighting intensity parameter λ is set to be 0 or 0.1, where the smoothing effect become weaker with larger λ . For each pre-smoothing scenario with (d, λ) , smoothed Wisepill adherence time series $Y_i = (Y_{i,1}, \dots, Y_{i,T})$ are clustered by the change-point mixture model, Fclust, distFPCA and k-centers FC approaches, respectively.

The performance of these clustering results is evaluated by comparing the original cluster membership vector $(c_1, \dots, c_n)$ with the clustering output membership $(\hat{c}_1, \dots, \hat{c}_n)$. We use adjusted Rand-Index(aRI) of Hubert and Arabie [13], which is a corrected version of the Rand index ([20]), to measure the similarity between clustering results and the 'true' clusters. The Rand index is calculated by $RI = \frac{a+c}{a+b+c+d}$ where a is the number of pairs of subjects that are clustered into same cluster from both clustering result, c is the number of pairs of subjects that are clustered into different cluster from both clustering result, and b and d are the number of pairs of subjects that are clustered into same cluster from one clustering result and different clusters from the other clustering result. Rand index is thus bounded between 0 and 1, where a larger number indicates better agreement between two

clustering results. aRI is the corrected-for-chance version of Rand index, with mean 0 and upper bound 1. Larger aRI indicates higher similarity between clusterings as well.

With each simulation scenario, 25 to 30 subject adherence patterns were generated for each of the three clusters. The aRI is calculated for the four clustering results for 100 repeated simulations. We present the mean and standard deviation of these aRI for comparison among different clustering approaches. Figure 2 illustrates an example simulation setting where the observed binary time series $W_{\lambda}(t)$ were generated from latent adherence curves $A_{\lambda}(t)$. $A_{\lambda}(t)$'s were simulated by three distinct change-point models with random noise. It is clear that with longer window length d the smoothed $Y_{\lambda}(t)$ has less variation within each cluster, which improves the separation of clusters. We found often the clustering results were quite poor when d is small (e.g., $d < 14$). Furthermore, small d may lead to non-convergence of one or more clustering algorithms, or the computation time becomes too long. Thus we used relative long d for the simulation results. We found that by fixing other parameters, $d = 28$ or $d = 56$ yielded negligible differences in clustering performance, thus we only present the results with $d = 28$. On the other hand, the choice of λ also produced different degrees of smoothness in Y_i , where larger λ yielded higher variations. However, varying λ had less pronounced impact on the clustering performance than different choice of d . Interestingly, the clustering performances from distFPCA, k-centers FC and change-point mixture model become worse with larger λ , but the performance of Fclust improves with larger λ (though its performance is still worse than the three other methods). This is likely due to the fact that Fclust is focused on modeling the covariance structure of spline coefficients, and larger variance of Y_i may contain more information for such structure that resulted in improved performance.

We found that when the change-point assumption is valid, results for scenario 1 and 2 in table 1 indicated that among non-parametric functional clustering, k-centers FC performed best with highest $mean(aRI)$. The parametric mixture model approach performed similarly to non-parametric approaches with small λ 's. As we assumed relative simple covariance structure for the change-point model, the mixture approach could be less robust against more complex variation in Y_i smoothed with larger λ . We also noticed that clustering with symmetrical window pre-smoothing yielded slightly better performance than that with historical window smoothing, for all clustering algorithms uniformly.

4.2 Performance Comparison, parametric mixture vs. semi-parametric

We have described the semi-parametric clustering approach in section 3.3. The semi-parametric approach has higher flexibility with the covariance structure over the parametric change-point mixture models. To quantify the degree of gain in clustering performance from additional complexity, two additional simulation studies were conducted. The performances of parametric mixture model and semi-parametric model were compared by their clustering accuracies versus the true cluster membership of simulated data. Data generation procedures, simulation design scenario parameters and pre-smoothing procedures were similar to those described in section 4.1. The true adherence process $A_{\lambda}(t)$ was generated by random multivariate normal distributions with change-point mean $f_{sim,\lambda}(t)|c_i = f_{c_i}(t|\theta^*)$ and covariance matrix Σ_{c_i} . The Σ_{c_i} 's were generated from (i). simple diagonal matrix: $\Sigma_{c_i} = \sigma_{c_i}^2 I$;

or (ii). complex covariance process expanded by fourier basis plus measurement error:

$$\Sigma_c = \sigma_c^2 I + \sum_{k=1}^{P_c} \lambda_k^c \Phi_k^c(s) \Phi_k^c(t), \text{ where } \lambda_k^c = 2^{-k}, p_c \in \{2, 3, 4\} \text{ and}$$

$$\Phi_k^c(t) = \sin(k\omega_c t + \frac{\pi}{2} \frac{1 + (-1)^k}{2}). \text{ These scenarios allow much more complex covariance}$$

structure that the simple assumption from change-point EM approach may be inadequate.

The binary irregular wisepill data generation and the preprocessing smoothing step were the same as in section 4.1. More simulation setup details can be found in supplementary attachment 1.

The parametric mixture and semi-parametric models were fitted to the true adherence processes $A(t)$'s as well as the smoothed process $Y(t)$'s. The reason for fitting these models to two different data sources was to demonstrate how the semi-parametric approach compares to parametric approach on analyzing intrinsically continuous versus pre-smoothed binary functional data. Table 2 summarizes the results from these simulations. We observe that the semi-parametric approach performed better for clustering $A(t)$'s, but did not demonstrate advantage in clustering $Y(t)$'s. Further assessment of the data generated from simulation revealed that the estimated covariance structure from $A(t)$'s and $Y(t)$'s were quite different. In fact, the estimated covariance surface of $A(t)$'s were similar to the true covariance surface used to generate data, but the estimated covariance surface of $Y(t)$'s were much closer to a diagonal surface after the discretization (generating binary $W(t)$) and pre-smoothing (obtaining $Y(t)$) steps. Plots of these estimated covariance surfaces are available in supplementary attachment 1. This explained why the change-point mixture model with simple covariance structure assumption performed well even when the true covariance structure is complex. Thus with the current pre-process procedure for binary observations from the wisepill measurement for PrEP adherence, the added flexibility of semi-parametric approach did not lead to improved clustering performance. In our application to the HPTN 069 study, we found that parametric and semiparametric change-point mixture models yielded similar results in terms of clustering.

5 Application to Wisepill data in HPTN 069 trial

We applied functional clustering methods to Wisepill-measured adherence monitoring data from the HPTN 069 trial. The study has enrolled 594 subjects randomly assigned to one of four active PrEP drug arms and were followed up for 1 year, 406 of whom were MSMs and 188 were female at risk. Each subject's Wisepill opening counts were collected daily over the 350 days follow-up, and were converted to yes/no binary records for each day. The "heartbeat" signals sent by the Wisepill device were used to verify data quality, because these indicate the monitoring device is in good working condition. We only use the "heartbeat" verified Wisepill daily opening data with at least 100 days long for the analysis, which were available from 219 study subjects. These binary time-series observations were often irregularly spaced in time due to various device technical issues such as cellular network compatibility, depleted battery, and other unknown reasons.

The clustering methods proposed above are available for irregular spaced functional data, thus we apply them directly to smoothed adherence patterns for the 182 study subjects with

relative complete Wisepill observations. Smoothed adherence were obtained with symmetrical smoothing windows with width of 57 days and constant weights for each day within the window, and we found that using different pre-smoothing tuning parameters (e.g., 28-day window, exponential weights) yield similar results.

We present the clustering result from both change-point mixture models and the k-centers FC. Figure 3 shows the result for total cluster number $C=3$. The left column is the result from the change-point mixture model, and the right column is the result from k-centers FC clustering. The colored solid curves in each panel are the local polynomial smoothed empirical average adherence levels for each cluster, and light grey curves are smoothed adherence levels from 10 randomly selected subjects in that cluster. For the change-point approach, the estimated cluster mean functions were also plotted in dotted colored curves in each corresponding panel. For the k-centers FC, the dotted colored curves were the estimated cluster centroid represented by Karhunen-Loeve expansion. We've re-ordered the cluster labels for the two columns to make the clusters in each row comparable.

The cluster in the top row (cluster 1) showed relatively high adherence according to Wisepill signals, and the adherence level remained relatively stable over time. The cluster in the middle row (cluster 2) had good starting adherence but the level continued to decrease over the follow up. The cluster in the bottom row (cluster 3) showed medium adherence at the beginning but the level decreased quickly and below 40% before 100th day. Qualitatively, the adherence pattern can be called “good”, “medium”, and “poor” adherence groups from top to bottom rows, and we notice the similarity of clusters in the same row by the two clustering methods. Since the k-centers FC is a non-parametric approach, the similarity of results from both k-centers FC and change-point mixture model provide reassuring evidence that the parametric assumption of linear change-point trend is plausible. The maximum likelihood estimates of model parameters ($\hat{\mu}_0, \hat{\tau}_0, \hat{\beta}_1, \hat{\beta}_2$) of each cluster were:

(0.90,200.0,0.00,0.00) in the “good” adherence cluster, (0.85,150.0, -0.0014, -0.0019) in the “medium” adherence cluster, and (0.57,82.1, -0.0038, -0.0020) in the “poor” adherence cluster. In the “good” adherence cluster, there does not appear to be a change-point as individuals generally maintain high adherence throughout the follow-up period. In the “medium” adherence cluster, individual's adherence drop over time, with an apparent accelerated drop after around 150 days. In the “poor” adherence cluster, changes in the rate of adherence decline happens after around 82 days. The change-point patterns are of great scientific interest and can be used to design effective intervention strategies to improve subjects' adherence in future HIV prevention studies.

The cluster assignment comparison for the two approaches are shown in table 3. The agreement of two clustering results was quite good with $RI=0.8628$ and $aRI=0.7054$. We found the main difference came from the 12 subjects assigned to cluster 1 by K-centers FC and cluster 2 by change-point mixture model. These subjects also help explain why the K-centers cluster 1 centroid showed more decreasing trend than the change-point mixture model cluster 1 mean estimate. The empirical means for clusters 2 and 3 were quite similar between the two approaches, which is also reflected in the comparison table.

After clustering of pill taking patterns, we fitted multivariate logistic regression models with subject baseline covariates as one type of subsequent analysis. Baseline variables collected in HPTN 069 include age, gender, race, marriage status, employment status, and educational level. We used stepwise AIC for selecting the variables for the final model. We fitted the subject cluster memberships from change-point latent class model and k-centers FC in two separate regression model selection procedures, and both selected education level and race as the predictors in the final model. With the “good adherence” cluster as reference, the odds ratio estimates and their corresponding 95% confidence intervals are listed in table 4. Despite the differences in estimated odds ratio values, results from both models indicate that higher education level and non-white race were correlated with worse adherence to PrEP drug prescription over time. One reviewer pointed out there could be potential improvement for clustering if incorporating covariates in the clustering models. We agree with the argument, as the so-called “1-step” approach that combines association analyses and clustering brings about addition benefit of reduced bias in estimates of association coefficients. The analyses we presented in this section however were based on the “3-step” approach, where clustering, membership assignment and association analyses were separated in serial steps. Detailed discussion of “1-step” versus “3-step” approaches is however beyond the scope of our paper, we refer readers to [23].

We’ve also fitted the Wisepill data with higher number of total clusters. The general finding is that both k-centers FC and change-point mixture model produce similar clustering results for up to $C=8$, and the non-parametric approach still indicated that the piece-wise linear parametric assumption fitted data well. Bayesian Information Criteria (BIC) is available for selection of C with the mixture model, and we generally found BIC decreasing with larger C . Although the within-cluster variation was reduced in each additional cluster, the average adherence level patterns were quite similar among some of the additional clusters. We did not present the results with larger C here because although the model fit was improved, there was not enough qualitative separation between additional clusters versus the previous clusters.

6 Discussion

We have adapted several parametric and nonparametric clustering methods for functional data for analyzing densely recorded discrete functional data, as they were previously developed for continuous functional data. We used two-step approaches, by first using a pre-smoothing process to transform raw data into continuous functional data and then cluster the smoothed data. Two types of clustering tools, parametric mixture models and non-parametric functional clustering methods, were considered and compared. The parametric mixture models were based on change-point cluster mean function assumption, motivated by prior scientific knowledge (from our scientific collaborators and also from medication adherence literature). For functional clustering, we compared methods suitable for non-regular functional data, including both a model-based approach that assumes random effect models with functional basis coefficients, and non-parametric distance/dissimilarity metric induced clustering algorithms with FPCA. Simulations showed that, when parametric assumption holds, the mixture model approach yielded similar or better performances than

functional approaches while being easily interpretable. On the other hand, non-parametric functional approaches work better when the parametric assumption is violated.

Additionally, we have proposed a semi-parametric model which was a semi-parametric clustering approach that combined model-based assumption for the mean process and functional estimation for unstructured covariance surfaces. Simulation results indicated that its improved flexibility could lead to better performance to handle data with complex covariance structures. However, in our particular example the covariance structure after the pre-processing step remained quite simple and the extra complexity in semi-parametric approach did not come up with better clustering performance.

We applied better performing clustering approaches to Wisepill data from the HPTN 069 study, and identified distinct clusters of pill box opening behaviour over time. The clusters found by the mixture models and the k-centers FC were very similar, and the temporal trend seemed to suggest that the change-point trends were plausible. The clustering results were then used as input for further analysis to identify personal characteristics that are associated with the Wisepill adherence behaviour.

The parametric mixture models have the advantage of interpretability, while non-parametric functional approaches are more flexible to adapt with more complex noise structures. The semi-parametric approach we proposed incorporates the advantages of these two type of approaches. On the other hand, careful investigation of the data structure should be performed to balance between potential gain in model performance and added complexity and computational burden, as illustrated in our simulation studies of change-point mixture versus semi-parametric approaches. The valid assumptions for (latent) raw data collected from mobile devices could be heavily altered by the (often necessary) pre-processing one uses. Another open problem in functional data analysis is when the missing at random mechanism does not hold, it is very difficult to use existing methods for irregular data. We explored several scenarios of missing mechanism in our simulation studies, and found that when a proportion of curves are right truncated, the functional clustering approaches gave poor results while parametric model performance was much better. With the working parametric assumption for the mean processes, the missing at random assumption is less violated. However this is not the case for non-parametric settings, because the missing proportion of curves were cross-extrapolated from other members sharing the cluster, which may have already been biased by informative missing. A third direction is to analyze discrete functional data directly without pre-smoothing. Hall et al [10] proposed a model to estimate the underlying latent Gaussian process of generalized longitudinal data, and Huang et al [12] applied their methods for clustering paired cocaine abuse treatment data. However, their works were developed for sparsely observed longitudinal data only. It would be of interest to generalize their work to densely recorded discrete functional data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This was partially supported by two grants from the National Institutes of Health (NIH), R01AI121259 and R01HL130483.

References

1. Akaike H: A new look at the statistical model identification. *IEEE transactions on automatic control* 19(6), 716–723 (1974)
2. Bélisle CJ: Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability* 29(4), 885–895 (1992)
3. Chiou JM, Li PL: Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(4), 679–699 (2007)
4. Fraley C, Raftery AE: Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* 97(458), 611–631 (2002)
5. Gable AR, Lagakos SW, et al.: *Methodological challenges in biomedical HIV prevention trials*. National Academies Press (2008)
6. Gibaldi M, Nagashima R, Levy G: Relationship between drug concentration in plasma or serum and amount of drug in the body. *Journal of Pharmaceutical Sciences* 58(2), 193–197 (1969)
7. Grant RM, Anderson PL, McMahan V, Liu A, Amico KR, Mehrotra M, Hosek S, Mosquera C, Casapia M, Montoya O, et al.: Uptake of pre-exposure prophylaxis, sexual practices, and hiv incidence in men and transgender women who have sex with men: a cohort study. *The Lancet infectious diseases* 14(9), 820–829 (2014) [PubMed: 25065857]
8. Gulick RM, Wilkin TJ, Chen YQ, Landovitz RJ, Amico KR, Young AM, Richardson P, Marzinke MA, Hendrix CW, Eshleman SH, et al.: Phase 2 study of the safety and tolerability of maraviroc-containing regimens to prevent hiv infection in men who have sex with men (hptn 069/actg a5305). *The Journal of infectious diseases* 215(2), 238–246 (2016)
9. Haberer JE, Kahane J, Kigozi I, Emenyonu N, Hunt P, Martin J, Bangsberg DR: Real-Time Adherence Monitoring for HIV Antiretroviral Therapy. *AIDS Behav* 14(6), 1340–1346 (2010) PMID: PMC2974938 [PubMed: 20809380]
10. Hall P, Müller HG, Yao F: Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(4), 703–723 (2008)
11. Hartigan JA, Wong MA: Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1), 100–108 (1979)
12. Huang H, Li Y, Guan Y: 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 (2014)
13. Hubert L, Arabie P: Comparing partitions. *Journal of classification* 2(1), 193–218 (1985)
14. Jacques J, Preda C: Functional data clustering: a survey. *Advances in Data Analysis and Classification* 8(3), 231–255 (2014)
15. James GM, Sugar CA: Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98(462), 397–408 (2003)
16. Kass RE, Raftery AE: Bayes factors. *Journal of the american statistical association* 90(430), 773–795 (1995)
17. Lewis AS, Overton ML: Nonsmooth optimization via quasi-newton methods. *Mathematical Programming* 141(1-2), 135–163 (2013)
18. Louis TA: Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 226–233 (1982)
19. Peng J, Muller HG: Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics* pp. 1056–1077 (2008)
20. Rand WM: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66(336), 846–850 (1971)

21. Schwarz G, et al.: Estimating the dimension of a model. *The annals of statistics* 6(2), 461–464 (1978)
22. Tibshirani R, Walther G: Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics* 14(3), 511–528 (2005)
23. Vermunt JK: Latent class modeling with covariates: Two improved three-step approaches. *Political analysis* 18(4), 450–469 (2010)
24. Vrijens B, Vincze G, Kristanto P, Urquhart J, Burnier M: Adherence to prescribed antihypertensive drug treatments: longitudinal study of electronically compiled dosing histories. *Bmj* 336(7653), 1114–1117 (2008) [PubMed: 18480115]
25. Ward JH Jr: Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58(301), 236–244 (1963)
26. Yao F, Muller HG, Wang JL, et al.: Functional linear regression analysis for longitudinal data. *The Annals of Statistics* 33(6), 2873–2903 (2005)

Fig. 1:

Example Wisepill adherence data from the HPTN 069 study, including both raw data (left) and smoothed adherence data (right). The orange/pink colored blocks in the raw data plot indicate open/no open Wisepill signal for subject (row) and time in days (column). Blank blocks indicate missing data. Each curve in the smoothed data plot was derived from the corresponding subject with same id in the raw data plot.

Fig. 2:

Simulated data from an example simulation run. In the top row, the left panel shows the simulated binary time series $W_{\lambda}(t)$'s, while right panel shows the corresponding latent adherence process $A_{\lambda}(t)$'s, with red, green and black referring to three clusters, respectively. The Middle rows shows estimated smoothed $Y_{\lambda}(t)$'s with constant weights (i.e., $\lambda = 0$) and smoothing window length $d = 14$ and $d = 28$ for the left and right panels, respectively. The bottom row shows estimated smoothed $Y_{\lambda}(t)$'s with exponentially decayed weights (i.e., $\lambda = 0.1$) and smoothing window length $d = 14$ and $d = 28$ for the left and right panels, respectively.

Fig. 3:
Estimated clusters of Wisepill adherence patterns from the HPTN 069 Study, change-point mixture (left column) vs. k-centers FC (right column). In each panel, the colored solid/dotted curves are the model estimated/empirical means from subjects in the cluster. Grey curves show smoothed Wisepill traces from 10 random subjects in each cluster.

Table 1:

Clustering results from 3 simulation scenarios. $A(t)$ followed change-point mean $f^*(t)$ assumption for scenario 1 and 2, and nonchange-point assumptions for scenario 3. C took value C for symmetric smoothing window and L for historical smoothing window. The window length d was chosen to be 28-29 days for reasonable smoothness to ensure good algorithm performances. Exponential weight intensity λ varied from 0 to 0.1 for different degree of smoothing strength.

			Fclust	distFPCA	k-centers FC	Chg-Pt EM
Scenario 1	$\lambda = 0$	aRI	.7408	.9291	.9401	.9108
Assume change-point mean		sd(aRI)	.2237	.0962	.0856	.0922
<hr/>						
$= L, d = 28$	$\lambda = 0.1$	aRI	.8027	.9186	.9371	.8915
		sd(aRI)	.2110	.1058	.0855	.1332
<hr/>						
Scenario 2	$\lambda = 0$	aRI	.7391	.9251	.9374	.9418
Assume change-point mean		sd(aRI)	.2243	.0974	.0865	.0850
<hr/>						
$= C, d = 29$	$\lambda = 0.1$	aRI	.8161	.9198	.9380	.9272
		sd(aRI)	.1940	.1089	.0848	.1136
<hr/>						
Scenario 3	$= L$	aRI	.5422	.6422	.6541	.0706
Assume non-change-point mean		sd(aRI)	.2640	.2232	.2304	.1689
<hr/>						
$d = 29, \lambda = 0$	$= C$	aRI	.5554	.6274	.6710	.1021
		sd(aRI)	.2404	.2259	.2111	.1939

Table 2:

Simulation comparing parametric mixture vs. semi-parametric. The clustering approaches were fitted to true adherence process $A(t)$ and smoothed wisepill process $Y(t)$ separately. The covariance structure for generating $A(t)$'s was based on either simple independent diagonal covariance matrix $\sigma_c^2 I$ or complex Fourier-basis

expanded structure $\Sigma_c = \sigma_c^2 I + \sum_{k=1}^{p_c} \lambda_k^c \Phi_k^c(s) \Phi_k^c(t)$. Performance of clustering were shown in adjusted Rand-index values versus the true cluster membership of subjects and their standard deviations: aRI(s.d.).

Input data	Σ_c structure	Chg-Pt EM	semi-parametric
$A(t)$	$\sigma_c^2 I$	.9677(.0684)	.9944(.0304)
	$\sigma_c^2 I + \sum_{k=1}^{p_c} \lambda_k^c \Phi_k^c(s) \Phi_k^c(t)$	.9676(.0387)	.9899(.0184)
$Y(t)$	$\sigma_c^2 I$	.9398(.0848)	.9124(.0936)
	$\sigma_c^2 I + \sum_{k=1}^{p_c} \lambda_k^c \Phi_k^c(s) \Phi_k^c(t)$	.8238(.1002)	.8030(.1128)

Table 3:

HPTN 069 Wisepill adherence cluster assignment comparison, change-point mixture model vs. K-centers FC.

		k-center FC			
		cluster 1	cluster 2	cluster 3	
Chg-Pt Mixture	cluster 1	82	0	0	82
	cluster 2	12	37	5	54
	cluster 3	1	3	42	46
		95	40	47	182

Table 4:

Application to the HPTN 069 Study: multivariate logistic regression of baseline covariates on Wisepill adherence clusters. Odds ratio estimates and their 95% confidence interval are shown, with the “good” adherence cluster as reference.

	Change-point mixture model		k-centers FC	
	medium	poor	medium	poor
Edu.(High school or less)				
College	1.14 (0.38, 3.37)	0.50 (0.18, 1.42)	0.67 (0.23, 1.90)	0.61 (0.22, 1.70)
Graduate school	0.26 (0.06, 1.17)	0.04 (0.004, 0.36)	0.12 (0.02, 0.69)	0.05 (0.005, 0.43)
Race (White)				
Black	3.19 (1.24, 8.22)	10.04 (3.69, 27.34)	3.70 (1.40, 9.75)	8.25 (3.22, 22.14)
Other	1.81 (0.70, 4.63)	3.51 (1.21, 10.16)	1.62 (0.59, 4.43)	2.32 (0.84, 6.43)