Electronic Journal of Statistics
Vol. 12 (2018) 171–193
ISSN: 1935-7524
https://doi.org/10.1214/17-EJS1389

Cluster analysis of longitudinal profiles with subgroups

Xiaolu Zhu and Annie Qu[*]

Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
e-mail: xzhu28@illinois.edu*;* anniequ@illinois.edu

Abstract: In this paper, we cluster profiles of longitudinal data using a penalized regression method. Specifically, we allow heterogeneous variation of longitudinal patterns for each subject, and utilize a pairwise-grouping penalization on coefficients of the nonparametric B-spline models to form subgroups. Consequently, we identify clusters based on different patterns of the predicted longitudinal curves. One advantage of the proposed method is that there is no need to pre-specify the number of clusters; instead the number of clusters is selected automatically through a model selection criterion. Our method is also applicable for unbalanced data where different subjects could have measurements at different time points. To implement the proposed method, we develop an alternating direction method of multipliers (ADMM) algorithm which has the desirable convergence property. In theory, we establish the consistency properties for approximated nonparametric function estimation and subgrouping memberships. In addition, we show that our method outperforms the existing competitive approaches in our simulation studies and real data example.

Keywords and phrases: ADMM, longitudinal data, minimax concave penalty, model selection, nonparametric spline method.

Received August 2016.

1. Introduction

In longitudinal data studies, distinguishing patterns of longitudinal trajectories is useful in many practical applications. For example, in personalized medicine, correctly identifying subgroups is essential for individualized treatment assignment, since distinguishing the dynamics of disease progression status among patients is critical in evaluating the effectiveness of a certain treatment. In time-course gene expression studies, grouping genes with similar expression profiles over time is also useful in association studies to identify crucial genes linked with certain diseases ([4], [14], [24]).

One way to distinguish longitudinal patterns is to apply cluster analysis through classical multivariate clustering methods and algorithms by treating repeated measurements as multivariate vectors. For example, the K-means method

[*] The authors would like to thank the editor, the associate editor and the referees for their insightful comments and suggestions.

[12] is one of the popular dissimilarity-measure-based approaches, which partitions subjects into a pre-specified number of clusters based on the Euclidean distance from each cluster mean. Alternatively, the Gaussian mixture model [11] assumes a finite mixture of multivariate normal distributions, and estimates the parameters of the distribution and the conditional membership probabilities. In addition, [25] provides a comprehensive review of several other applicable multivariate clustering algorithms.

However, there are several drawbacks to treating longitudinal data as multivariate vectors, since this assumes that the multivariate vector is balanced with the same number of time points. That is, the multivariate clustering approach cannot be applied directly unless the missing entries are imputed first. Most critically, the multivariate-vector method does not take the information of time ordering into account, and consequently, the clustering result is invariant to different permutations of measurements within subjects, which makes little sense when observations are over time and the trajectory patterns are of our main interest. Alternatively, to capture the growth curves of each subject, we can cluster the trajectories of subjects with nonparametric smoothing approaches, such as B-spline techniques. For example, [1] partitions subjects through the K-means method using B-spline coefficients. [17, 18] and [7] apply spline approximations under the linear mixed-effects model framework. However, the imposed parametric assumption of the mixed effects model, typically a normal distribution, makes their methods less flexible in practice.

In addition, the aforementioned clustering approaches require one to specify the number of clusters in advance, which could be problematic in cluster analysis. Recent developments in penalized regression methods allow one to estimate the cluster centers and select the number of clusters simultaneously. [20] models the multivariate vectors assuming an individual center for each subject and penalizes the pairwise distance between two subjects' centers. [5] utilizes a convex penalty and considers the clustering as a convex optimization problem. [19] incorporates covariates of interest to model univariate response data, which assumes different intercepts for different subjects. However, none of these approaches models trajectories over time for longitudinal data.

In this paper, we propose a regression-based approach which partitions observations into subgroups through penalization of pairwise distances between the B-spline coefficients vectors. One advantage of the proposed approach is that a pre-specification of the number of clusters is not required; instead, we select the number of clusters automatically through a model selection criterion. This allows us to achieve model estimation and subgrouping subjects simultaneously. Another advantage is that the proposed method is applicable in characterizing longitudinal trajectories which can include unbalanced longitudinal data. In addition, we implement an alternating directions and method of multipliers algorithm (ADMM) [2] to achieve fast convergence of the method. In theory, we establish the consistency property for the proposed method which can identify the true underlying subgroup membership asymptotically. Furthermore, our simulation studies and real data analysis also confirm that, compared to other existing approaches, the proposed method performs well in identifying subgroups.

The rest of the article is organized as follows. Section 2 introduces the model formulation. In Section 3, we propose a nonparametric pairwise-grouping approach along with the ADMM algorithm, and establish theoretical properties and implementation strategies. Simulation studies and real data analysis are presented in Sections 4 and 5. We conclude the article with a brief discussion in Section 6.

2. A subject-wise model for longitudinal data

The subject-wise model for subject i $(i = 1, \cdots, n)$ is formulated as follows:

$$y_{ij} = f_i(x_{ij}) + \varepsilon_{ij}, \tag{2.1}$$

where y_{ij} is the response at the jth $(j = 1, \cdots, n_i)$ repeated measurement and x_{ij} is a one-dimensional covariate which defines the pattern of interest, and the random errors ε_{ij} are uncorrelated with mean 0 and variance σ^2. In this paper, we only deal with the setting where x_{ij} is a covariate of time, and investigate the patterns of longitudinal trajectories over time. In principle, we can extend our method to a more general setting with more than one covariate. Without loss of generality, we assume that the covariates x_{ij} can be scaled to a compact interval $\mathcal{X} = [0, 1]$. For this subject-wise model, each subject has its unique unknown smoothing function denoted as $f_i(\cdot) \in C^q(\mathcal{X})$, which is assumed to be qth-order continuously differentiable.

The estimation of the subject-wise smoothing functions $f_i(\cdot)$ characterizing the longitudinal profile of a specific covariate is one of our main interests. Here, we estimate f_i by the nonparametric B-spline approach, which flexibly approximates smoothing functions. We define the qth-order B-splines with a set of m internal knots sequences $\boldsymbol{\kappa} = \{0 = \kappa_0 < \kappa_1 < \cdots < \kappa_m < \kappa_{m+1} = 1\}$ recursively [8] as

$$B_l^1(x) = \begin{cases} 1, & \kappa_l \leq x < \kappa_{l+1} \\ 0, & \text{o.w.} \end{cases},$$

$$\text{and } B_l^q(x) = \frac{x - \kappa_l}{\kappa_{l+q-1} - \kappa_l} B_l^{q-1}(x) + \frac{\kappa_{l+q} - x}{\kappa_{l+q} - \kappa_{l+1}} B_{l+1}^{q-1}(x).$$

Then $f_i(x)$ can be approximated as $f_i(x) \approx s_i(x) = \sum_l B_l^q(x)\beta_{il} = B(x)^T\boldsymbol{\beta}_i$ through a linear combination of B-spline bases, where $B(x)$ is a B-spline basis vector and $\boldsymbol{\beta}_i$ is a p-dimensional coefficient vector with $p = m + q$, determined by the number of knots m and B-spline order q.

Let $\mathbf{f}_i = (f_i(x_{i1}), \cdots, f_i(x_{in_i}))^T$, $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})^T$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots \varepsilon_{in_i})^T$. The matrix form of the model (2.1) can be reformulated as

$$\mathbf{Y} = \mathbf{f} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y} = \left(\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T\right)^T$, $\mathbf{f} = \left(\mathbf{f}_1^T, \cdots, \mathbf{f}_n^T\right)^T$, and $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_1^T, \cdots, \boldsymbol{\varepsilon}_n^T\right)^T$. In addition, the corresponding B-spline approximation is

$$\mathbf{f} \approx \mathbf{s} = \mathbf{B}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_n^T)^T$, $\mathbf{s} = (\mathbf{s}_1^T, \cdots, \mathbf{s}_n^T)^T$, $\mathbf{s}_i = B_i\boldsymbol{\beta}_i$, $\mathbf{B} = \text{diag}(B_1, \cdots, B_n)$, and $B_i = (B(x_{i1}), \cdots, B(x_{in_i}))^T$ is an $n_i \times p$ basis matrix for the subject i.

We assume that the subjects share the same smoothing function form if they are from the same group. That is, $f_i = f_j$ if subjects i and j are from the same cluster group. Consequently, let $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_K\}$ be a partition of $\{1, \cdots, n\}$, where $K(K \leq n)$ is the number of distinct groups. We define the nonparametric function subspace $\mathcal{M}_{\mathcal{G}}^{\mathbf{f}}$ corresponding to the group partition as

$$\mathcal{M}_{\mathcal{G}}^{\mathbf{f}} = \left\{\mathbf{f} : f_i = f_{(k)}, f_i \in C^q(\mathcal{X}), \text{for any } i \in \mathcal{G}_k, 1 \leq k \leq K\right\},$$

and the subspace of the B-spline coefficients corresponding to the group partition as

$$\mathcal{M}_{\mathcal{G}}^{\boldsymbol{\beta}} = \left\{\boldsymbol{\beta} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_{(k)}, \boldsymbol{\beta}_i \in \mathbf{R}^p, \text{for any } i \in \mathcal{G}_k, 1 \leq k \leq K\right\}.$$

Our goal is to identify the distinct group patterns of the smoothing functions for any given subjects. This is equivalent to distinguishing between B-spline coefficients for each group.

3. Methodology and theory

3.1. A nonparametric pairwise-grouping approach

In this subsection, we propose a pairwise-grouping approach through penalization to achieve B-spline coefficient estimation and grouping of subjects simultaneously. We adopt a penalized B-spline approach [9] to utilize a relatively large number of knots, but impose a penalty on the B-spline coefficients. More specifically, the objective function of the penalized regression spline given the dth-order difference penalty is

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\left\|\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}\right\|_2^2 + \frac{1}{2}\lambda_1\boldsymbol{\beta}^T\mathbf{D}_d\boldsymbol{\beta} = \frac{1}{2}\sum_{i=1}^{n}\left\{\left\|\mathbf{y}_i - B_i\boldsymbol{\beta}_i\right\|_2^2 + \lambda_1\boldsymbol{\beta}_i^T D_d\boldsymbol{\beta}_i\right\}, \tag{3.1}$$

where $\|\cdot\|_2$ is an L_2 norm, $\mathbf{D}_d = \text{diag}\left(D_d, \cdots, D_d\right)$, $D_d = \Delta_d^T\Delta_d$ and Δ_d is a $(p-d)\times p$ matrix presentation of the dth-order difference operator. For example, the second-order difference operator Δ_2 is defined as

$$\Delta_2 = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

The penalized B-spline coefficient estimator is obtained by minimizing the following objective function (3.1):

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathcal{M}^{\boldsymbol{\beta}}} Q(\boldsymbol{\beta}) = \left(\mathbf{B}^T\mathbf{B} + \lambda_1\mathbf{D}_d\right)^{-1}\mathbf{B}^T\mathbf{Y},$$

where $\mathcal{M}^{\boldsymbol{\beta}} = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^{np}\}$. Consequently, the estimation of the smoothing function approximation is $\tilde{\mathbf{f}} = \mathbf{B}\tilde{\boldsymbol{\beta}}$. Notice that this is equivalent to applying the penalized B-spline approach for each subject separately under the subject-wise model framework. In the penalized spline approach, [22] provides a rule-of-thumb to select about $\min\{n_i/4, 35\}$ number of knots for subject i, where the location of knots can be chosen based on sample quantiles of $\{x_i\}$.

To identify subgroups corresponding to distinct smoothing functions, we group subjects together if they possess similar functional forms of nonparametric approximations. Specifically, we penalize pairwise distances of B-spline coefficients to encourage subjects to fall into the same group. We propose the corresponding objective function as:

$$L(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + \sum_{i,j\in\mathcal{L}} \rho\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \lambda_2\right), \tag{3.2}$$

where $\rho(\cdot, \lambda_2)$ is a penalty function with a tuning parameter λ_2, and $\mathcal{L} = \{l = (i,j) : 1 \le i < j \le n\}$ is the index set containing the total number of possible subject pairs $|\mathcal{L}| = n(n-1)/2$.

The essence of the proposed approach is to take advantage of the flexibility of nonparametric approximation while controlling the complexity of the model, which is also associated with the number of subgroups. Here, the tuning parameter λ_2 plays such a role to determine the number of subgroups. By minimizing the objective function (3.2), we simultaneously obtain nonparametric coefficient estimation and group subjects if their estimated nonparametric coefficient vectors are sufficiently close.

In general, the choice of penalty function $\rho(\cdot, \lambda_2)$ is critical since it results in different parameter estimation and subgroup selection. For example, a Lasso-type of penalty leads to a sparse solution, which could be appealing in merging subjects into groups. However, it is also well-known that the Lasso estimation is biased. Here, we apply the minimax concave penalty (MCP) [29], which is nearly unbiased and also has the sparsity property. Specifically, the penalty function is

$$\rho\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \lambda_2\right) = \rho_\tau\left(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2, \lambda_2\right),$$

where $\rho_\tau(t,\gamma) = \gamma\int_0^t (1 - \frac{x}{\tau\gamma})_+ dx$, and the regularization parameter τ controls the unbiasedness and concavity of the penalty function. We obtain $\hat{\boldsymbol{\beta}}$ through minimizing (3.2) using the MCP penalty, and the corresponding smoothing function estimation is $\hat{\mathbf{f}} = \mathbf{B}\hat{\boldsymbol{\beta}}$. With this nearly-unbiasedness property, we can achieve accurate nonparametric coefficient estimation and group membership recovery.

In fact, it is challenging to optimize the objective function (3.2) directly, as the proposed grouping penalty is not separable in terms of $\boldsymbol{\beta}_i$'s. Here, we develop an alternative approach which introduces a new set of parameters $\mathbf{v}_l = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j, l \in \mathcal{L}$, which are equivalent to the pairwise differences of B-spline coefficient vectors. Consequently, the above optimization problem can be transformed into the following constrained problem:

$$\begin{aligned} &\min Q(\boldsymbol{\beta}) + \sum_{l\in\mathcal{L}} \rho_\tau(\|\mathbf{v}_l\|_2, \lambda_2), \\ &\text{subject} \quad \text{to} \quad \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \mathbf{v}_l = \mathbf{0}. \end{aligned}$$

We solve the above optimization problem using the alternating direction method of multipliers (ADMM), which is a variant of the augmented Lagrange multipliers (ALM) method. The above constrained problem can be further converted to an optimization problem through augmenting a quadratic penalty with a fixed parameter θ:

$$\begin{aligned} &\min Q(\boldsymbol{\beta}) + \sum_{l\in\mathcal{L}} \rho_\tau(\|\mathbf{v}_l\|_2, \lambda_2) + \frac{\theta}{2}\sum_{l\in\mathcal{L}} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \mathbf{v}_l\right\|_2^2, \\ &\text{subject} \quad \text{to} \quad \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \mathbf{v}_l = \mathbf{0}. \end{aligned}$$

Notice that the above two problems are equivalent due to the fact that the imposed quadratic penalty is zero for any feasible $\boldsymbol{\beta}$ and $\mathbf{v} = (\mathbf{v}_1^T, \cdots, \mathbf{v}_{|\mathcal{L}|}^T)^T$ satisfying the constraint. Therefore, we can estimate parameters by minimizing the corresponding Lagrangian as follows:

$$\begin{aligned} L_\theta(\boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\lambda}) &= Q(\boldsymbol{\beta}) + \sum_{l\in\mathcal{L}} \rho_\tau(\|\mathbf{v}_l\|_2, \lambda_2) \\ &+ \frac{\theta}{2}\sum_{l\in\mathcal{L}} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \mathbf{v}_l\right\|_2^2 + \sum_{l\in\mathcal{L}} \boldsymbol{\lambda}_l^T(\mathbf{v}_l - \boldsymbol{\beta}_i + \boldsymbol{\beta}_j), \end{aligned}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \cdots, \boldsymbol{\lambda}_{|\mathcal{L}|}^T)^T$ are Lagrange multipliers.

Following the AMDD algorithm, we update the estimates of $\boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\lambda}$ sequentially at the $(s+1)$th iteration step as follows:

$$\boldsymbol{\beta}^{s+1} = \arg\min_{\boldsymbol{\beta}} L_\theta(\boldsymbol{\beta}, \mathbf{v}^s, \boldsymbol{\lambda}^s), \tag{3.3}$$

$$\mathbf{v}^{s+1} = \arg\min_{\mathbf{v}} L_\theta(\boldsymbol{\beta}^{s+1}, \mathbf{v}, \boldsymbol{\lambda}^s), \tag{3.4}$$

$$\boldsymbol{\lambda}_l^{s+1} = \boldsymbol{\lambda}_l^s + \theta(\mathbf{v}_l^{s+1} - \boldsymbol{\beta}_i^{s+1} + \boldsymbol{\beta}_j^{s+1}), l \in \mathcal{L}.$$

Note that the first minimization function in Eq. (3.3) is equivalent to minimizing a quadratic function:

$$f(\boldsymbol{\beta}) = \frac{1}{2}\left\|\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}\right\|_2^2 + \frac{\lambda_1}{2}\boldsymbol{\beta}^T\mathbf{D}_d\boldsymbol{\beta} + \frac{\theta}{2}\sum_{l\in\mathcal{L}} \left\|\tilde{\mathbf{v}}_l - A_l\boldsymbol{\beta}\right\|_2^2,$$

where $\tilde{\mathbf{v}}_l = \mathbf{v}_l + \frac{1}{\theta}\boldsymbol{\lambda}_l$ and $A_l = (\mathbf{e}_i - \mathbf{e}_j)^T \otimes I_p$, in which $\otimes$ is the Kronecker product and $\mathbf{e}_i$ is an n-dimensional vector with one at the ith component and zeros otherwise. Let $\mathbf{A} = (A_1^T, \cdots, A_{|\mathcal{L}|}^T)^T$, and $\tilde{\mathbf{v}} = (\tilde{\mathbf{v}}_1^T, \cdots, \tilde{\mathbf{v}}_{|\mathcal{L}|}^T)^T$, then we have $\boldsymbol{\beta}^{s+1} = \left(\mathbf{B}^T\mathbf{B} + \lambda_1\mathbf{D}_d + \theta\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{B}^T\mathbf{Y} + \theta\mathbf{A}^T\tilde{\mathbf{v}}^s\right)$, where $\tilde{\mathbf{v}}^s = \mathbf{v}^s + \frac{1}{\theta}\boldsymbol{\lambda}^s$ is estimated in the previous iteration.

As for the second minimization function in Eq. (3.4), it is a convex function with respect to each $\mathbf{v}_l$ if $\tau > 1/\theta$, even though $L_\theta(\boldsymbol{\beta}^{s+1}, \mathbf{v}, \boldsymbol{\lambda}^s)$ involves a non-convex MCP penalty term. Consequently, we can update $\mathbf{v}_l^{s+1}$ explicitly as

$$\mathbf{v}_l^{s+1} = \begin{cases} \mathbf{u}_l^{s+1} & \text{if} \quad \|\mathbf{u}_l^{s+1}\|_2 \geq \tau\lambda_2, \\ \frac{\tau\theta}{\tau\theta - 1}(1 - \frac{\sigma}{\|\mathbf{u}_l^{s+1}\|_2})_+ \mathbf{u}_l^{s+1} & \text{if} \quad \|\mathbf{u}_l^{s+1}\|_2 < \tau\lambda_2, \end{cases}$$

where $\sigma = \lambda_2/\theta$ and $\mathbf{u}_l^{s+1} = \boldsymbol{\beta}_i^{s+1} - \boldsymbol{\beta}_j^{s+1} - \boldsymbol{\lambda}_l^s/\theta$.

In non-convex optimization, it is important to assign appropriate initial values to obtain a good solution. We choose to initialize the ADMM algorithm with a warm start which also reduces the number of iterations. Specifically, we first apply the penalized B-spline for each subject and assign $\boldsymbol{\beta}^0 = \tilde{\boldsymbol{\beta}} = \left(\mathbf{B}^T\mathbf{B} + \lambda_1\mathbf{D}_d\right)^{-1}\mathbf{B}^T\mathbf{Y}$, where λ_1 can be selected using the BIC_{λ_1} from the two-step tuning procedure in section 3.3.

The detailed ADMM algorithm is outlined as follows:

Algorithm 1 ADMM algorithm

```
Initialize:
    $\boldsymbol{\lambda}^0 = \mathbf{0}$ and $\boldsymbol{\beta}^0$, $\mathbf{v}^0 = \arg\min_{\mathbf{v}} L_\theta(\boldsymbol{\beta}^0, \mathbf{v}, \boldsymbol{\lambda}^0)$, $\theta$ and $\tau > \frac{1}{\theta}$ are fixed.
for $s = 0, 1, 2, \cdots$ do
    $\boldsymbol{\beta}^{s+1} = \left(\mathbf{B}^T\mathbf{B} + \lambda_1\mathbf{D}_d + \theta\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{B}^T\mathbf{Y} + \theta\mathbf{A}^T\tilde{\mathbf{v}}^s\right)$, where $\tilde{\mathbf{v}}^s = \mathbf{v}^s + \frac{1}{\theta}\boldsymbol{\lambda}^s$,
    $\mathbf{v}^{s+1} = \arg\min_{\mathbf{v}} L_\theta(\boldsymbol{\beta}^{s+1}, \mathbf{v}, \boldsymbol{\lambda}^s)$,
    $\boldsymbol{\lambda}_l^{s+1} = \boldsymbol{\lambda}_l^s + \theta(\mathbf{v}_l^{s+1} - \boldsymbol{\beta}_i^{s+1} + \boldsymbol{\beta}_j^{s+1})$, for all $l \in \mathcal{L}$.
    if stopping criteria are met then
        break
    end if
end for
```

In the above algorithm, the stopping criteria are evaluated based on $\mathbf{r}_l^{s+1} = \boldsymbol{\beta}_i^{s+1} - \boldsymbol{\beta}_j^{s+1} - \mathbf{v}_l^{s+1}$ and $\mathbf{d}_k^{s+1} = -\theta\left(\sum_{i=k}(\mathbf{v}_l^{s+1} - \mathbf{v}_l^s) - \sum_{j=k}(\mathbf{v}_l^{s+1} - \mathbf{v}_l^s)\right)$. Define $\mathbf{r} = \left(\mathbf{r}_1^T, \cdots, \mathbf{r}_{|\mathcal{L}|}^T\right)^T$ and $\mathbf{d} = \left(\mathbf{d}_1^T, \cdots, \mathbf{d}_n^T\right)^T$. The algorithm terminates at the step s^* if $\|\mathbf{r}^{s^*}\|_2 \leq \epsilon^r$ and $\|\mathbf{d}^{s^*}\|_2 \leq \epsilon^d$, where ϵ^r and ϵ^d are small numbers according to [2]:

$$\epsilon^d = \sqrt{np}\epsilon^{abs} + \epsilon^{rel}\|\mathbf{A}^t\boldsymbol{\lambda}^{s*}\|_2, \qquad \epsilon^r = \sqrt{|\mathcal{L}|p}\epsilon^{abs} + \epsilon^{rel}\max\left\{\|\mathbf{A}\boldsymbol{\beta}^{s*}\|_2, \|\mathbf{v}^{s*}\|_2\right\},$$

where the parameters ϵ^{abs} and ϵ^{rel} are predetermined small values.

Theorem 3.1. *The above ADMM algorithm converges, such that* $\left\|\mathbf{r}^{s+1}\right\|_2^2 \to 0$ *and* $\left\|\mathbf{d}^{s+1}\right\|_2^2 \to 0$ *for a sufficiently large iteration step* s.

Theorem 3.1 establishes the convergence property of the ADMM algorithm, indicating that the stopping criteria of the algorithm can be reached as the number of iteration increases. The proof of Theorem 3.1 is provided in the appendix A.1.

3.2. Asymptotic properties

In this subsection, we establish the asymptotic properties of the estimators obtained by the proposed approach. To study the convergence rate of $\hat{\mathbf{f}}$, we first provide some regularity conditions.

(C1). Suppose that the design points $\{x_{ij}\}_{i=1,j=1}^{n,n_i}$ follow a density function f_X, which is absolutely continuous, and there exist constants c_1 and c_2 such that $0 < c_1 \le \min_{x\in\mathcal{X}} f_X(x) \le \max_{x\in\mathcal{X}} f_X(x) \le c_2 < \infty$.

(C2). The error terms in model (2.1) are uncorrelated with a mean zero and a variance $\sigma^2 > 0$.

(C3). For each f_i $(i = 1, \cdots, n)$, $f_i \in C^q(\mathcal{X})$ is a q-th order continuously differentiable function defined on a compact set $\mathcal{X} = [0, 1]$.

(C4). The set of knots is defined as $\boldsymbol{\kappa}^m = \{0 = \kappa_0 < \kappa_1 < \cdots < \kappa_m < \kappa_{m+1} = 1\}$. Let $\delta = \max_{0\le l\le m}(\kappa_{l+1} - \kappa_l)$, there exists a constant $M > 0$ such that $\delta / \min_{0\le l\le m}(\kappa_{l+1} - \kappa_l) \le M$, and $\delta = o(m^{-1})$.

(C5). The number of knots $m = o(n_0)$, where $n_0 = \min(n_1, \cdots, n_n)$.

(C6). Assume $N_k = O(N)$, where $N_k = \sum_{i\in\mathcal{G}_k} n_i$ for $k = 1, \cdots, K$, and $N = \sum_{i=1}^n n_i$.

Conditions (C1) - (C5) are standard assumptions for nonparametric B-spline smoothing functions. Similar conditions are also given by [6], [30], and [26]. In (C5), the condition on the number of knots applies for all subjects. In addition, we impose a constraint on cluster size in (C6), implying that the cluster size grows as the sample size increases.

We first investigate the convergence property on the estimation of the penalized B-spline approximation $\tilde{\mathbf{f}}$. Let $\mathbf{f}^o$ be the true function corresponding to the true group partition $\mathcal{G}$. We establish the estimation consistency in the following Lemma 3.1.

Lemma 3.1. *Under conditions (C1) - (C5), for any fixed* n*, and given a sufficiently large* n_0*, such that* $\gamma_d = (p-d)(\frac{\lambda_1 \tilde{c}}{n_0})^{1/2d} < 1$*, where* $\tilde{c} = c\{1 + o(1)\}$*, and* $c = \pi^{2d}\left(\int_0^1 f_X(x)^{1/2d}dx\right)^{-2d}$*. We establish the following convergence rate of* $\tilde{\mathbf{f}}$*:*

$$\left\|\tilde{\mathbf{f}} - \mathbf{f}^o\right\|_n^2 \le O_p(\frac{m}{n_0}) + O_p(\frac{\lambda_1^2}{n_0^2}m^{2d}) + O_p(m^{-2q}).$$

Lemma 3.1 shows that the convergence rate of the penalized spline estimator is determined by three factors. The first term is the average asymptotic variance, which decreases as the number of repeated measurements grows and increases if the nonparametric model is more complex with an increasing number of knots. The second term is introduced by the shrinkage bias, but vanishes if $\lambda_1 \to 0$. The last term reflects the nonparametric approximation bias, which is also related to the model complexity. We show in Lemma 3.1 that the convergence rate also depends on the minimum number of repeated measurements n_0 among subjects. The proof of Lemma 3.1 is given in the appendix A.2.

When the true group membership is known, we obtain the oracle approximation by $\tilde{\mathbf{f}}^{or} = \mathbf{B}\tilde{\boldsymbol{\beta}}^{or}$, where the corresponding oracle penalized spline estimator is

$$\tilde{\boldsymbol{\beta}}^{or} = \arg \min_{\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}^{\boldsymbol{\beta}}} Q(\boldsymbol{\beta}).$$

Let $N_0 = \min(N_1, \cdots, N_K)$, we provide the convergence rate of the oracle approximation in the following Lemma 3.2.

Lemma 3.2. *Under conditions (C1) - (C4) and (C6), given a sufficiently large N_0, such that $\gamma_d = (p-d)(\frac{\lambda_1 \tilde{c}}{N_0})^{1/2d} < 1$, we have*

$$\left\|\tilde{\mathbf{f}}^{or} - \mathbf{f}^o\right\|_n^2 = O_p(\frac{m}{N}) + O_p(\frac{\lambda_1^2}{N^2} m^{2d}) + O_p(m^{-2q}).$$

In contrast to the convergence rate in Lemma 3.1 for the penalized B-spline estimators, Lemma 3.2 establishes a faster convergence rate for the oracle penalized spline estimators when the true subgroup information is known, since $N > n_0$. The above convergence property is guaranteed as long as the number of repeated measurements for each cluster is sufficiently large, as it is equivalent to obtaining $\tilde{\boldsymbol{\beta}}^{or}$ within each subgroup. The proof of Lemma 3.2 is provided in the appendix A.2.

In the following, let b be the minimum distance between smoothing functions $\mathbf{f}_{(k)}^o$ and $\mathbf{f}_{(k')}^o$ from any two clusters, that is, $b = \min_{k \neq k'} \|\mathbf{f}_{(k)}^o - \mathbf{f}_{(k')}^o\|$. We denote the proposed approximation as $\hat{\mathbf{f}} = \mathbf{B}\hat{\boldsymbol{\beta}}$.

Theorem 3.2. *Under conditions (C1) - (C6), and if $cb \geq \tau\lambda_2$ holds for a constant $c > 0$, then for any fixed n, and given a sufficiently large n_0, such that $\gamma_d = (p-d)(\frac{\lambda_1 \tilde{c}}{n_0})^{1/2d} < 1$, we have*

$$\left\|\hat{\mathbf{f}} - \mathbf{f}^o\right\|_n^2 = O_p(\frac{m}{n_0}) + O_p(\frac{\lambda_1^2}{n_0^2} m^{2d}) + O_p(m^{-2q}).$$

Theorem 3.2 indicates that the convergence rate of the proposed approximation is the same as the penalized spline estimators as long as there is a sufficiently large number of repeated measurements for each subject. In addition, the distance between smoothing functions from any two clusters should be sufficiently large to achieve the above convergence rate. The details of the proof are given in the appendix A.3.

Corollary 3.1 (Subgroup membership recovery consistency)**.** *Suppose the regularity conditions in Theorem 3.2 hold, then the subgroup memberships satisfy* *the following properties: for any* $i, j \in \mathcal{G}_k$, $\left\|\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j\right\|_n^2 \leq O_p(m^{-2q})$, *and for any* $i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'$, $\left\|\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j\right\|_n^2 \geq b^2 - O_p(m^{-2q})$ *as* $n_0 \to \infty$.

Corollary 3.1 indicates that the recovery of subgroup membership depends on both the minimum distance b and the nonparametric approximation bias $O_p(m^{-2q})$. If the true functions from two clusters are very close to each other, then the nonparametric approximation needs to be sufficiently accurate with little bias. Otherwise, a smoother function with fewer knots is adequate to ensure subgroup membership recovery as long as the distance between two clusters is sufficiently large. The proof is given in the section A.3.

3.3. Implementation

In this section, we provide the strategy of forming subjects into subgroups based on parameter estimation. In addition, we provide the tuning parameters selection criteria.

In fact, we form clusters based on estimated parameters $\hat{\mathbf{v}}$ instead of B-spline coefficients $\hat{\boldsymbol{\beta}}$. In the proposed approach, we encourage closeness between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ through pairwise grouping penalization. That is, subjects i and j are expected to be clustered in the same group if $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_j$. However, this is not achievable since we impose a quadratic penalty on $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \mathbf{v}_l$ in the ADMM algorithm when augmenting the optimization constraint $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j = \mathbf{v}_l$, and this makes the implementation problematic for clustering. Therefore, we apply the MCP penalty to $\hat{\mathbf{v}}_l$ to enjoy the sparsity property in the algorithm, and merge subjects i and j into the same cluster if $\hat{\mathbf{v}}_l = 0$.

To select the tuning parameters λ_1 and λ_2, we propose a two-step procedure to search from a sequence of grid points. Note that λ_1 controls the smoothness of the B-spline approximation, and λ_2 controls the number of clusters selected, denoted as $\hat{K}$. Traditionally, we can implement a grid search for both tuning parameters simultaneously. However, this increases the computational cost tremendously. To solve this problem, we propose a two-step procedure which first searches for an optimal value of λ_1 given $\lambda_2 = 0$, then selects λ_2 given the optimal λ_1 from the first step. Although this procedure may not lead to the optimal selection for both tuning parameters, our numerical studies show that this strategy works effectively. More specifically, we select λ_1 by minimizing

$$BIC_{\lambda_1} = \sum_{i=1}^{n} \left\{ \log\left(\frac{\left\|\mathbf{y}_i - \hat{\mathbf{f}}_i\right\|_2^2}{n_i} \right) + \frac{\log(n_i)}{n_i} \mathrm{df}_i \right\},$$

where $\mathrm{df}_i = tr\left\{\mathbf{B}_i(\mathbf{B}_i^T\mathbf{B}_i + \lambda_1 D_d)^{-1}\mathbf{B}_i^T\right\}$, and λ_2 is selected by minimizing

$$BIC_{\lambda_2} = \log\left(\frac{\left\|\mathbf{Y} - \hat{\mathbf{f}}\right\|_2^2}{n} \right) + \frac{\log(n) * \mathrm{df}}{n}, \text{where } \mathrm{df} = \frac{\hat{K}}{n}\sum_{i=1}^{n} \mathrm{df}_i.$$

4. Simulation study

In this section, we conduct simulation studies to investigate the performance of the proposed nonparametric pairwise-grouping approach (NPG) when the subjects have unbalanced numbers of repeated measurements, which often arises in practice. We compare the proposed method with the smoothing spline regression clustering approach [18], using the original unbalanced data. However, traditional multivariate-vector approaches are not feasible for handling unbalanced data unless imputation for missing entries is implemented. Instead, we compare our method to the K-means (bKmeans) and the Gaussian Mixtures (bGM) methods by treating the subject-wise penalized B-spline estimators $\tilde{\boldsymbol{\beta}}_i$'s as multivariate vectors.

We implement the K-means method with R function *kmeans* and select the number of clusters based on the Gap statistic [13] using the R package *cluster*. To ensure the robustness of the K-means method, we calculate a mean result from 10 random picks of initial centers. The Gaussian mixtures approach is implemented by the R package *mclust*, and the number of clusters is selected based on the embedded Bayesian Information Criterion (BIC), which is chosen from $K = 1, 2, \cdots, 15$ in each simulation. We also implement the smoothing spline regression clustering approach with the R package *MFDA*. In addition, we compare the results with a mixture of mixed-effects method (Mixed) with P-spline smoothing technique [7]. In our simulation, we fix $\theta = 1$ and $\tau = 2$ to ensure the convexity of our objective function. For the methods we compare, the proper tuning parameters are chosen accordingly for each approach. The final results are based on 100 simulations. We implement the proposed approach in R software, and the programming codes are available on Github (https://github.com/Xiaolu-Zhu/LongitudinalClustering.git).

To evaluate the performance of these clustering algorithms, we calculate the estimated number of groups $\hat{K}$ selected and several frequently used external validity measures: the Rand index [21], the adjusted Rand index (aRand) [15] and the Jaccard index [16]. Let the true positive (TP) be the number of pairs of subjects from the same cluster and assigned to the same cluster, the true negative (TN) be the number of pairs of subjects from different clusters and assigned to different clusters, the false positive (FP) be the number of pairs of subjects from different clusters but assigned to the same cluster, and the false negative (FN) be the number of pairs of subjects from the same cluster but assigned to different clusters. The Rand index is calculated as Rand $= \frac{\text{TP+TN}}{\text{TP + TN + FP + FN}}$, which measures the percentage of pairwise agreements between the true and selected clusters. However, Rand tends to be large even under random partitions. The adjusted Rand (aRand) index corrects this problem, and is calculated by $\frac{\text{Rand - E(Rand)}}{\text{max(Rand) - E(Rand)}}$. The Jaccard index is calculated as $\frac{\text{TP}}{\text{TP+ FN +FP}}$. For these external criteria, a higher value indicates a better agreement between the selected and the true group memberships.

4.1. Case 1: Independent measurement

In this simulation setting, we generate 15 subjects from each of the following four distinct functional patterns: $f_{(1)}(x) = \cos(2\pi x)$; $f_{(2)}(x) = 1 - 2\exp(-6x)$; $f_{(3)}(x) = -1.5x$; and $f_{(4)}(x) = -1.5x + 1.5$. The continuous response y_{ij} for subject i from the kth subgroup is generated by $y_{ij} = f_{(k)}(x_{ij}) + \varepsilon_{ij}$, $k = 1, \cdots, 4$, $j = 1, \cdots, 10$, where random errors within subjects are independent $\varepsilon_{ij} \sim^{iid} N(0, 0.4^2)$, and $\{x_{ij}\}_{j=1}^{10}$ are equally spaced points on $[0, 1]$. In order to mimic real data situations, we allow 30% of the subjects from each subgroup to have 40% missing repeated measurements. The number of knots is recommended by [22] as $\min\{n_i/4, 40\}$ for subject i. We choose the B-spline with an order $q = 3$ and the number of knots $m = 3$ for all subjects.

Table 1 shows that the proposed approach performs the best in terms of three external criteria. Since the K-means and Gaussian mixtures methods approximate each subject's pattern individually, the clustering results are not as good as the proposed method. In addition, the proposed NPG method is able to identify the true function subgroups more effectively, as it estimates the B-spline coefficients for all subjects simultaneously and borrows cross-subject information from the same subgroup. Our simulations show that the mixed-effects model performs better than the K-means and Gaussian mixtures methods, but slightly worse than the proposed approach. We also compare the computational cost among these methods. In this simulation setting, the average computational time for the "Mixed" method in [7] is 23.13 seconds, the K-means method is 12.53 seconds, the Gaussian mixture method is 1.06 seconds and the MFDA method is 12.37 seconds, while the proposed NPG approach with fixed tuning parameters takes about 27.18 seconds.

Table 1

Comparison results from the proposed nonparametric pairwise-grouping (NPG), K-means (bKmeans), Gaussian Mixtures (bGM), MFDA and mixed-effects (Mixed) method.

Methods	$\hat{K}$	**Rand**	**aRand**	**Jaccard**
NPG	4.01	0.995	0.986	0.980
bKmeans	1.34	0.317	0.091	0.296
bGM	6.82	0.914	0.745	0.673
MFDA	5.99	0.941	0.823	0.755
Mixed	3.73	0.959	0.904	0.886

As shown in Corollary 3.1, subgroup membership can be recovered successfully regardless of the number of knots used in B-spline approximation as long as the underlying longitudinal function patterns are far from each other. We illustrate this using various numbers of knots from 1 to 3. Figure 1 and Figure 2 indicate that the underlying true mean function curves are recovered well even with one knot, although the non-linear curves can be approximated more accurately if we increase the number of knots to 3, especially for the nonlinear

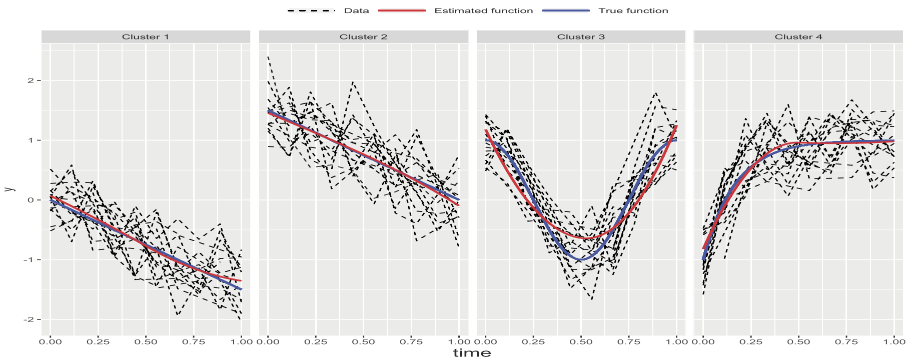

FIG 1. *B-spline function recovery and subgrouping in Case 1 when the number of knots is 1.*

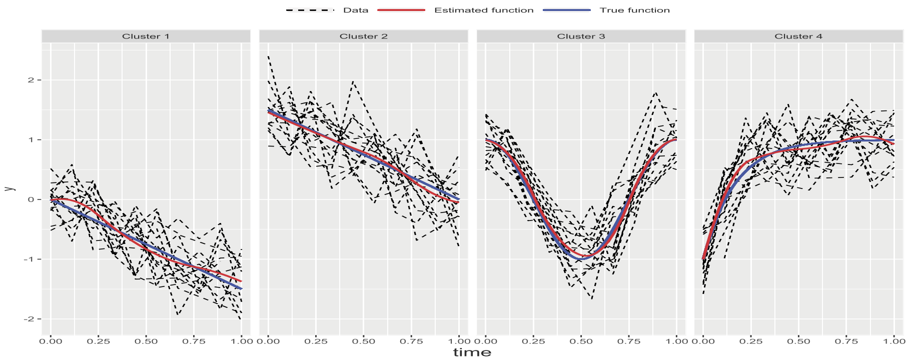

FIG 2. *B-spline function recovery and subgrouping in Case 1 when the number of knots is 3.*

patterns in clusters 3 and 4. However, using one or three knots leads to the same accurate subgrouping identification in this simulation setting.

We also conduct simulation experiments to investigate the effect of tuning parameter selection. Specifically, we fix the tuning parameter $\theta = 1$ and $\tau = 2$ to ensure that $\tau > \frac{1}{\theta}$ is satisfied. We compare the NPG performance on a range of $\tau \in (2, 4, 8, 16)$, which shows that the clustering results are quite stable with the Rand index range in (0.993, 0.995), the adjusted Rand index in (0.982, 0.986) and the Jaccard index in (0.975, 0.980).

The clustering performance is also compared under an unbalanced data setting, where we follow the same data generation process, except that the sample sizes for the four subgroups are 5, 10, 20 and 40, respectively. Table 2 shows similar clustering accuracy as in the balanced data setting, where both the proposed NPG and Mixed models show robust performance as reflected in the three different indices.

Table 2

Comparison results from the proposed nonparametric pairwise-grouping (NPG), K-means (bKmeans), Gaussian Mixtures (bGM), MFDA and mixed-effects (Mixed) method for unbalanced data setting.

Methods	$\hat{K}$	**Rand**	**aRand**	**Jaccard**
NPG	4.02	0.990	0.978	0.974
bKmeans	1.13	0.392	0.033	0.385
bGM	5.57	0.848	0.650	0.615
MFDA	5.98	0.838	0.615	0.563
Mixed	3.67	0.966	0.930	0.923

4.2. Case 2: Correlated measurement

In this simulation study, we investigate the performance of the proposed approach when the repeated measurements are correlated. In particular, we generate data from the same process as in the Case 1 balanced setting, but allow random errors to have a certain correlation structure. Specifically, we generate the true $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, 0.4^2\mathbf{R}_0)$, where $\mathbf{R}_0$ is either AR(1) or exchangeable (EX) with a 0.7 correlation coefficient.

Since the original NPG approach is based on the independence assumption of errors from the same subject, we modify the proposed approach by utilizing the working correlation structure. The corresponding objective function becomes

$$L_R(\boldsymbol{\beta}) = \frac{1}{2}\big(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}\big)^T \mathbf{R}^{-1}\big(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}\big) + \frac{1}{2}\lambda_1 \boldsymbol{\beta}^T \mathbf{D}_d \boldsymbol{\beta} + \sum_{i,j \in \mathcal{L}} \rho\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \lambda_2\right),$$

where $\mathbf{R}$ is a block diagonal matrix with diagonal components of a given working correlation matrix, such as an AR(1) or exchangeable correlation structure from each subject. Notice that the modified NPG method with an independent working correlation matrix is equivalent to the original NPG, where $\mathbf{R}$ is the identity matrix. In practice, we can estimate the working correlation coefficient using the residuals obtained from the NPG approach. We denote the modified NPG methods as NPGex and NPGar, corresponding to the exchangeable and AR(1) working correlation structures, respectively.

Tables 3 and 4 indicate that the proposed method performs similarly using either AR(1) or exchangeable working correlation structures, but outperforms other methods in distinguishing different functional patterns. In addition, we also calculate the average mean square error (AMSE) of the predictions of outcomes in Tables 3 and 4, showing that the misspecification of error correlation structure may lead to slight loss of estimation efficiency. In fact, [28] also mentions that we can gain improvement by utilizing a correct correlation structure when the correlation coefficient is large. However, this does not affect accuracy very much in identifying true group membership in this simulation setting.

Table 3

Comparison results from the proposed nonparametric pairwise-grouping (NPG), the modified nonparametric pairwise-grouping with AR(1) working correlation (NPGar), the modified nonparametric pairwise-grouping with exchangable working correlation (NPGex), K-means (bKmeans), Gaussian Mixtures (bGM), MFDA and mixed-effects (Mixed) methods when the true correlation structure is AR(1).

Methods	$\hat{K}$	**Rand**	**aRand**	**Jaccard**	**AMSE**
NPG	4.10	0.983	0.953	0.934	1.510
NPGar	4.05	0.986	0.960	0.943	1.493
NPGex	4.05	0.986	0.960	0.943	1.494
bKmeans	2.48	0.577	0.393	0.491	-
bGM	5.55	0.948	0.847	0.794	-
MFDA	5.97	0.928	0.781	0.706	-
Mixed	3.17	0.879	0.715	0.666	-

Table 4

Comparison results from the proposed nonparametric pairwise-grouping (NPG), the modified nonparametric pairwise-grouping with AR(1) working correlation (NPGar), the modified nonparametric pairwise-grouping with exchangeable working correlation (NPGex), K-means (bKmeans), Gaussian Mixtures (bGM), MFDA and mixed-effects (Mixed) methods when the true correlation structure is exchangeable.

Methods	$\hat{K}$	**Rand**	**aRand**	**Jaccard**	**AMSE**
NPG	4.48	0.976	0.933	0.906	1.522
NPGar	4.22	0.983	0.953	0.933	1.500
NPGex	4.18	0.986	0.961	0.944	1.470
bKmeans	2.43	0.529	0.332	0.447	-
bGM	5.93	0.922	0.771	0.706	-
MFDA	5.86	0.926	0.777	0.703	-
Mixed	3.04	0.876	0.712	0.661	-

5. An application to IRI marketing data

In this section, we apply the longitudinal clustering methods on an IRI marketing dataset. This dataset was developed by the SymphonyIRI Group [3], and consists of 11-year (2001-2011) weekly sales data of packaged goods from chain grocery and drug stores in 47 markets across the country. To better capture the overall sales pattern for each product category for 11 years at the market level, we aggregate the data from the weekly level to the yearly level from store specific to geographic market hierarchy, and from granularly itemized products to product category level. In this analysis, we focus on the Los Angeles market, where we aim to cluster 26 packaged goods categories into subgroups based on their longitudinal sales units trajectories, where the trajectories are standardized to remove the mean and control the unit variance.

We compare the clustering results using the proposed method (NPG), the mixed-effects (Mixed) method, the K-means and Gaussian mixture methods on B-spline coefficients fitted at subject level. The MFDA method is excluded from comparison because of too much computation instability to be deliverable. The

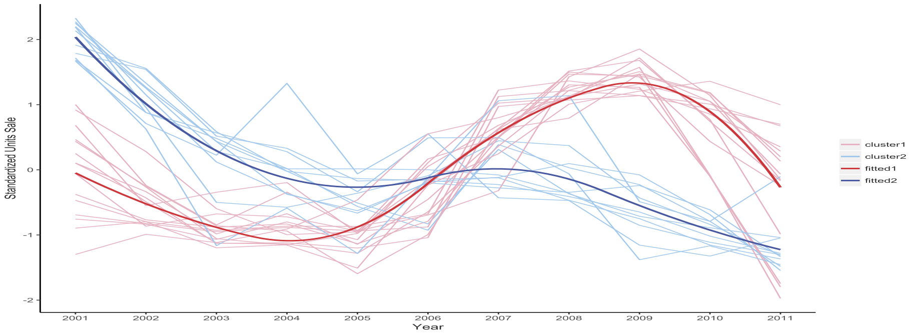

Fig 3. *The NPG method clustering results with fitted curves.*

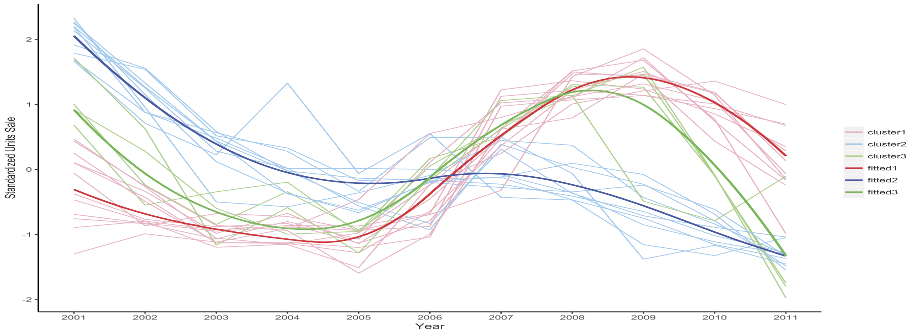

Fig 4. *The Mixed method clustering results with fitted curves.*

NPG approach utilizes 3 knots with an order of 3 for B-spline approximation and the Mixed approach specifies 5 knots based on its implementation guidelines.

Figures 3 and Figure 4 provide the clustering results and the fitted functional curves for the NPG and Mixed approaches. Specifically, the NPG method is able to identify two subgroups of trajectories, where cluster 1 identifies food related packaged products: beer/ale/alcoholic, cider, coffee, cold cereal, frozen dinners/entrees, frozen pizza, hotdogs, mayonnaise, milk, mustard/ketchup, peanut butter, salty snacks, soup, spaghetti/Italian sauce, sugar substitutes, and yogurt. Cluster 2 identifies non-food related products: blades, cigarettes, deodorant, diapers, facial tissue, paper towels, photography supplies, razors, shampoo, toilet tissue, and toothpaste. On the other hand, the Mixed method separates the product categories into 3 clusters, where cluster 3 has food products including mayonnaise, milk, and mustard/ketchup, and the non-food product toothpaste. The remaining two clusters are subsets of the two clusters from the NPG method. In addition, the fitted curves in Figures 3 and 4 for two clusters have similar patterns for these two methods.

The Gaussian Mixture model fails to identify any informative subgroups with one single cluster, while the K-means method requires random initialization. We implement 10 random picks of initial centers for the subject-wise fitted B-spline coefficient vectors, and the number of selected clusters from K-means ranges from 2 to 5, which leads to the same clustering result as the NPG method when the Gap statistic selects 2 clusters given a certain random initialization.

6. Discussion

In this article, we propose a nonparametric pairwise-grouping approach to cluster longitudinal trajectories over time. The new approach captures underlying functional patterns through utilizing the nonparametric B-spline method. In addition, we subgroup subjects through penalizing pairwise distances of B-spline coefficient vectors, which borrows between-subject information to better recover the true functions. The proposed NPG approach has the advantage of avoiding overfitting, compared to existing methods which approximate the underlying functions separately. This strategy works effectively when some of the repeated measurements are missing.

The proposed approach takes advantage of the MCP penalty, which is nearly unbiased and also leads to a sparse solution. This is especially important as we select the optimal tuning parameters through a model selection criterion BIC, which relies on model estimation accuracy. Note that other non-convex penalty functions, such as SCAD [10] or TLP [23], can also be applied here to utilize the unbiasedness property. However, the implementation and convergence property of the ADMM algorithm based on other viable penalty functions may require further investigation.

In this paper, although we assume an independence structure of random errors within subjects, a modified approach utilizing working correlation is also proposed to account for the correlation information. We show in simulation studies that the modified approach has similar performance in clustering as the NPG approach assuming independence, but leads to improved efficiency in estimation. The theoretical properties of the modified NPG method need to be further investigated if correlation information is of our interest.

The proposed approach can also be extended to subgroup identification incorporating multiple covariates under generalized linear models. One potential research topic is to extend the proposed framework for binary longitudinal outcomes, and to identify the subgroups of treatment effects. Identifying subgroups for binary data could be quite challenging, as specifying the proper loss function while taking the correlation within subjects into account is nontrivial.

Appendix A: Proofs

A.1. Proof of Theorem 3.1

Proof. Let $h(\boldsymbol{\beta}) = \frac{1}{2}\left\|\mathbf{Y}-\mathbf{B}\boldsymbol{\beta}\right\|_2^2+\frac{\lambda_1}{2}\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}$, $g(\mathbf{v}) = \sum_{l\in\mathcal{L}} g(\mathbf{v}_l) = \sum_{l\in\mathcal{L}} \rho_\tau(\|\mathbf{v}_l\|_2, \lambda_2)$, and $m(\boldsymbol{\beta}, \mathbf{v}) = \frac{\theta}{2}\sum_{l\in\mathcal{L}} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \mathbf{v}_l\right\|_2^2$, we write the Lagrangian $L_\theta(\boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\lambda}) =$

$h(\boldsymbol{\beta})+g(\mathbf{v})+m(\boldsymbol{\beta},\mathbf{v})-\boldsymbol{\lambda}^T\left(\mathbf{A}\boldsymbol{\beta}-\mathbf{v}\right)$. Then there exist $\boldsymbol{\lambda}^*$ such that $L_\theta(\boldsymbol{\beta},\mathbf{v},\boldsymbol{\lambda})$ is minimized and $\mathbf{A}\boldsymbol{\beta}^*-\mathbf{v}^*=0$, which implies that $L_\theta(\boldsymbol{\beta}^*,\mathbf{v}^*,\boldsymbol{\lambda}^*)\le L_\theta(\boldsymbol{\beta},\mathbf{v},\boldsymbol{\lambda}^*)$ holds for any other $(\boldsymbol{\beta},\mathbf{v})$. Write $p^*=h(\boldsymbol{\beta}^*)+g(\mathbf{v}^*)$, we have

$$p^*\le p^{s+1}+\frac{\theta}{2}\left\|\mathbf{r}^{s+1}\right\|_2^2-(\boldsymbol{\lambda}^*)^T\mathbf{r}^{s+1}, \tag{A.1}$$

where $p^{s+1}=h(\boldsymbol{\beta}^{s+1})+g(\mathbf{v}^{s+1})$ and $\mathbf{A}\boldsymbol{\beta}^{s+1}-\mathbf{v}^{s+1}=\mathbf{r}^{s+1}$.

Let $L_\theta^{\boldsymbol{\beta}}(\boldsymbol{\beta},\mathbf{v},\boldsymbol{\lambda})=h(\boldsymbol{\beta})+\frac{\theta}{2}\left\|\mathbf{A}\boldsymbol{\beta}-\tilde{\mathbf{v}}\right\|_2^2$ and $L_\theta^{\mathbf{v}_l}(\boldsymbol{\beta},\mathbf{v},\boldsymbol{\lambda})=\frac{1}{2}\left\|\mathbf{v}_l-\mathbf{u}_l\right\|_2^2+\frac{1}{\theta}g(\mathbf{v}_l)$ which are convex w.r.t $\boldsymbol{\beta}$ and $\mathbf{v}_l$ accordingly, where $\mathbf{u}_l=\boldsymbol{\beta}_i-\boldsymbol{\beta}_j-\frac{1}{\theta}\boldsymbol{\lambda}_l$. As $\boldsymbol{\lambda}_l^{s+1}=\boldsymbol{\lambda}_l^s-\theta\mathbf{r}_l^{s+1}$, it is straightforward that $\tilde{\mathbf{v}}^s=\mathbf{v}^s+\frac{1}{\theta}(\boldsymbol{\lambda}^{s+1}+\theta\mathbf{r}^{s+1})$.

By the definition, $\boldsymbol{\beta}^{s+1}$ minimizes $L_\theta^{\boldsymbol{\beta}}(\boldsymbol{\beta},\mathbf{v}^s,\boldsymbol{\lambda}^s)$ and we have that

$$\begin{aligned}\mathbf{0} &\in \partial L_\theta^{\boldsymbol{\beta}}(\boldsymbol{\beta}^{s+1},\mathbf{v}^s,\boldsymbol{\lambda}^s)=\partial h(\boldsymbol{\beta}^{s+1})+\theta(\mathbf{A}\boldsymbol{\beta}^{s+1}-\tilde{\mathbf{v}}^s)^T\mathbf{A}\\ &= \partial h(\boldsymbol{\beta}^{s+1})+\theta(\mathbf{v}^{s+1}-\mathbf{v}^s-\frac{1}{\theta}\boldsymbol{\lambda}^{s+1})^T\mathbf{A},\end{aligned}$$

which implies that

$$h(\boldsymbol{\beta}^{s+1})+\left(\theta(\mathbf{v}^{s+1}-\mathbf{v}^s)-\boldsymbol{\lambda}^{s+1}\right)^T\mathbf{A}\boldsymbol{\beta}^{s+1}\le h(\boldsymbol{\beta}^*)+\left(\theta(\mathbf{v}^{s+1}-\mathbf{v}^s)-\boldsymbol{\lambda}^{s+1}\right)^T\mathbf{A}\boldsymbol{\beta}^*.$$

Similarly, the following holds:

$$g(\mathbf{v}^{s+1})+\left(\boldsymbol{\lambda}^{s+1}\right)^T\mathbf{v}^{s+1}\le g(\mathbf{v}^*)+\left(\boldsymbol{\lambda}^{s+1}\right)^T\mathbf{v}^*.$$

By some arrangement and simplification, we have that

$$p^{s+1}-p^*\le(\boldsymbol{\lambda}^{s+1})^T\mathbf{r}^{s+1}-\theta\left(\mathbf{v}^{s+1}-\mathbf{v}^s\right)^T\left(\mathbf{r}^{s+1}+\mathbf{v}^{s+1}-\mathbf{v}^*\right). \tag{A.2}$$

Adding equations (A.1) and (A.2) together, we show that

$$\mathbf{0}\le(\boldsymbol{\lambda}^{s+1}-\boldsymbol{\lambda}^*)^T\mathbf{r}^{s+1}-\theta\left(\mathbf{v}^{s+1}-\mathbf{v}^s\right)^T\left(\mathbf{r}^{s+1}+\mathbf{v}^{s+1}-\mathbf{v}^*\right)+\frac{\theta}{2}\left\|\mathbf{r}^{s+1}\right\|_2^2.$$

Define $V^s=\frac{1}{\theta}\left\|\boldsymbol{\lambda}^s-\boldsymbol{\lambda}^*\right\|_2^2+\theta\left\|\mathbf{v}^s-\mathbf{v}^*\right\|_2^2$, we can show that $V^s-V^{s+1}\ge\theta\left\|\mathbf{v}^{s+1}-\mathbf{v}^s\right\|_2^2$, and that V^s decreases in each iteration. Therefore, we have $\sum_{s=0}^{\infty}\left(\theta\left\|\mathbf{v}^{s+1}-\mathbf{v}^s\right\|_2^2\right)\le V^0$, which implies that $\left\|\mathbf{v}^{s+1}-\mathbf{v}^s\right\|_2^2\to 0$. Results for $\left\|\mathbf{r}^{s+1}\right\|_2^2\to 0$ can be shown similarly as in [19] which is omitted here. □

A.2. Proof of Lemmas 3.1–3.2

Let $\|\cdot\|_2$ be the usual L_2 norm for functions or vectors. Let $L_2(\mathcal{X})$ be the space of all square integrable functions on $\mathcal{X}=[0,1]$, then $\left\|f\right\|_2^2=\int_0^1 f(x)^2dx$ for any

$f \in L_2(\mathcal{X})$. We define the theoretical and empirical norms as $\|f\|^2 = E[f(X)^2]$ and $\big\|f\big\|_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n} f(X_i)^2$, where $X_i's$ are a random sample of size n on $\mathcal{X}$. Let B be the orthonormal B-spline basis and its corresponding smoothing function be $s(x) = B(x)\boldsymbol{\beta}$.

Lemma A.1. *Under condition (C1), there exist constants $C \geq c > 0$, such that for any $f \in L_2(\mathcal{X})$, we have $c\|f\|_2 \leq \|f\| \leq C\|f\|_2$.*

Proof. This proof is straightforward with the definition of norms and the condition (C1) on the density of the design points. □

Lemma A.2. *Let $\boldsymbol{\beta}$ be any p-dimensional vector, there exist constants $C \geq c > 0$, such that $c\big\|\boldsymbol{\beta}\big\|_2^2 \leq \|B\boldsymbol{\beta}\|^2 \leq C\big\|\boldsymbol{\beta}\big\|_2^2$.*

Proof. This result follows from Lemma A.1. □

Lemma A.3. *There exist constants $C \geq c > 0$, such that except in an event whose probability tends to zero as $n \to \infty$, $c\|s\|^2 \leq \big\|s\big\|_n^2 \leq C\|s\|^2$ for any smoothing function s.*

Proof. The proof follows similarly to the proof of Lemma 4 in [27]. □

Lemma A.4. *There exist constants $C \geq c > 0$, such that except in an event whose probability tends to zero as $n \to \infty$, $c\big\|\boldsymbol{\beta}\big\|_2^2 \leq \big\|B\boldsymbol{\beta}\big\|_n^2 \leq C\big\|\boldsymbol{\beta}\big\|_2^2$ for any p-dimensional vector $\boldsymbol{\beta}$.*

Proof. The result follows from Lemma A.2 and Lemma A.3. □

Proof of Lemma 3.1. Notice that it is equivalent to individually obtaining $\tilde{\boldsymbol{\beta}}_i = \arg\min\limits_{\boldsymbol{\beta}_i} Q_i(\boldsymbol{\beta}_i)$, where $Q_i(\boldsymbol{\beta}_i) = \frac{1}{2}\big\|\mathbf{y}_i - B_i\boldsymbol{\beta}_i\big\|_2^2 + \frac{1}{2}\lambda_1\boldsymbol{\beta}_i^T D_d\boldsymbol{\beta}_i$. Therefore, the approximation of the smoothing function for subject i is $\tilde{\mathbf{f}}_i = B_i\tilde{\boldsymbol{\beta}}_i$. According to [6], we have $\big\|\tilde{\mathbf{f}}_i - \mathbf{f}_i^o\big\|_n^2 = O_p(\frac{m}{n_i}) + O_p(\frac{\lambda_1^2}{n_i^2}m^{2d}) + O_p(m^{-2q})$ when $\gamma_d < 1$. Consequently, we show that $\big\|\tilde{\mathbf{f}}_i - \mathbf{f}_i^o\big\|_n^2 \leq O_p(\frac{m}{n_0}) + O_p(\frac{\lambda_1^2}{n_0^2}m^{2d}) + O_p(m^{-2q})$ for all $i = 1, \cdots, n$. Thus it can be shown that, for fixed n,

$$\big\|\tilde{\mathbf{f}} - \mathbf{f}^o\big\|_n^2 = \frac{1}{N}\sum_{i=1}^{n}(\tilde{\mathbf{f}}_i - \mathbf{f}_i^o)^T(\tilde{\mathbf{f}}_i - \mathbf{f}_i^o) \leq O_p(\frac{m}{n_0}) + O_p(\frac{\lambda_1^2}{n_0^2}m^{2d}) + O_p(m^{-2q}).$$

□

Proof of Lemma 3.2. When the true group membership is known, it is equivalent to estimating $\tilde{\boldsymbol{\beta}}_{(k)}^{or} = \arg\min\limits_{\boldsymbol{\beta}} Q_{(k)}(\boldsymbol{\beta})$, where $Q_{(k)}(\boldsymbol{\beta}) = \sum\limits_{i\in\mathcal{G}_k}\big\{\big\|\mathbf{y}_i - B_i\boldsymbol{\beta}\big\|_2^2 + \lambda\boldsymbol{\beta}^T D_d\boldsymbol{\beta}\big\}$. Let $\tilde{\mathbf{f}}_{(k)}^{or} = (B_i)_{i\in\mathcal{G}_k}\tilde{\boldsymbol{\beta}}_k^{or}$ be the estimated functions belonging to the kth group and correspondingly $\mathbf{f}_{(k)}^{o} = (\mathbf{f}_i^o)_{i\in\mathcal{G}_k}$ be the true functions in the kth group. According to [6], we have $\left\|\tilde{\mathbf{f}}_{(k)}^{or}-\mathbf{f}_{(k)}^{o}\right\|_n^2=\frac{1}{N_k}\sum\limits_{i\in\mathcal{G}_k}(\tilde{\mathbf{f}}_i^{or}-\mathbf{f}_i^o)^T(\tilde{\mathbf{f}}_i^{or}-\mathbf{f}_i^o)=O_p(\frac{m}{N_k})+O_p(\frac{\lambda_1^2}{N_k^2}m^{2d})+O_p(m^{-2q})$ when $\gamma_d<1$. Thus it can be shown that

$$\left\|\tilde{\mathbf{f}}^{or}-\mathbf{f}^{o}\right\|_n^2=\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\tilde{\mathbf{f}}_i^{or}-\mathbf{f}_i^o)^T(\tilde{\mathbf{f}}_i^{or}-\mathbf{f}_i^o)=O_p(\frac{m}{N})+O_p(\frac{\lambda_1^2}{N^2}m^{2d})+O_p(m^{-2q}).$$

□

A.3. Proofs of Theorem 3.2 and Corollary 3.1

Proof of Theorem 3.2. By the triangular inequality, $\left\|\hat{\mathbf{f}}-\mathbf{f}^o\right\|_n^2\le\left\|\hat{\mathbf{f}}-\tilde{\mathbf{f}}^{or}\right\|_n^2+\left\|\tilde{\mathbf{f}}^{or}-\mathbf{f}^o\right\|_n^2$. If we can show that $\left\|\hat{\mathbf{f}}-\tilde{\mathbf{f}}^{or}\right\|_n^2=\left\|\mathbf{B}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2=O_p(\frac{m}{n_0})+O_p(\frac{\lambda_1^2}{n_0^2}m^{2d})+O_p(m^{-2q})$, then we can prove the theorem together with Lemma 3.2. That is, we aim to show that for a sufficiently large n_0 and any $\epsilon>0$, there exists a sufficiently large constant C such that

$$P\left(\inf_{\left\|\mathbf{B}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2=C(\frac{m}{n_0}+\frac{\lambda_1^2}{n_0^2}m^{2d}+m^{-2q})}L(\boldsymbol{\beta})>L(\tilde{\boldsymbol{\beta}}_{or})\right)\ge1-\epsilon.\tag{A.3}$$

This implies that there exists a local minimum of $L(\boldsymbol{\beta})$ which lies in the ball $\mathcal{B}=\{\boldsymbol{\beta}:\left\|\mathbf{B}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2\le C(\frac{m}{n_0}+\frac{\lambda_1^2}{n_0^2}m^{2d}+m^{-2q})\}$. As a result, $\left\|\mathbf{B}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2=O_p(\frac{m}{n_0}+\frac{\lambda_1^2}{n_0^2}m^{2d}+m^{-2q})$.

From Lemmas 3.1 and 3.2, we have $\left\|\tilde{\mathbf{f}}-\mathbf{f}^o\right\|_n^2\le O_p(\frac{m}{n_0})+O_p(\frac{\lambda_1^2}{n_0^2}m^{2d})+O_p(m^{-2q})$ and $\left\|\tilde{\mathbf{f}}^{or}-\mathbf{f}^o\right\|_n^2=O_p(\frac{m}{N})+O_p(\frac{\lambda_1^2}{N^2}m^{2d})+O_p(m^{-2q})$, which indicates that $\left\|\mathbf{B}(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2=\left\|\tilde{\mathbf{f}}-\tilde{\mathbf{f}}^{or}\right\|_n^2\le O_p(\frac{m}{n_0})+O_p(\frac{\lambda_1^2}{n_0^2}m^{2d})+O_p(m^{-2q})$ by the triangular inequality. Thus we have $\left\|\mathbf{B}(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2\le C_1(\frac{m}{n_0}+\frac{\lambda_1^2}{n_0^2}m^{2d}+m^{-2q})$ for a constant C_1.

As $\|\mathbf{B}(\tilde{\boldsymbol{\beta}}_{(k)}^{or}-\tilde{\boldsymbol{\beta}}_{(k')}^{or})\|_n\ge\|\mathbf{f}_{(k)}^o-\mathbf{f}_{(k')}^o\|_n-\|\mathbf{f}_{(k)}^o-\tilde{\mathbf{f}}_{(k)}^{or}\|_n-\|\mathbf{f}_{(k')}^o-\tilde{\mathbf{f}}_{(k')}^{or}\|_n$, we have $\|\mathbf{B}(\tilde{\boldsymbol{\beta}}_{(k)}^{or}-\tilde{\boldsymbol{\beta}}_{(k')}^{or})\|_n\ge b$ for a sufficiently large N. Lemma A.4 entails that there exists a constant c, $\|\tilde{\boldsymbol{\beta}}_{(k)}^{or}-\tilde{\boldsymbol{\beta}}_{(k')}^{or}\|_2\ge cb$. Similarly, for a sufficiently large n_0, we have $\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_2\ge cb$, for any $i\in\mathcal{G}_k$, $j\in\mathcal{G}_{k'}$ and $\boldsymbol{\beta}$ such that $\left\|\mathbf{B}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}^{or})\right\|_n^2=C(\frac{m}{n_0}+\frac{\lambda_1^2m^{2d}}{n_0^2}+m^{-2d})$.

Let $P_{\lambda_2}(\boldsymbol{\beta})=\sum\limits_{i,j\in\mathcal{L}}\rho_\tau\left(\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_2,\lambda_2\right)$. To show Eq. (A.3), using $\rho_\tau(0,\lambda_2)=0$ and $\rho_\tau(\cdot,\lambda_2)\ge0$, we have $P_{\lambda_2}(\tilde{\boldsymbol{\beta}}^{or})=\sum\limits_{i\in\mathcal{G}_k,j\in\mathcal{G}_{k'},k\ne k'}\rho_\tau\left(\|\tilde{\boldsymbol{\beta}}_i^{or}-\tilde{\boldsymbol{\beta}}_j^{or}\|_2,\lambda_2\right)$. Therefore

$$L(\boldsymbol{\beta})-L(\tilde{\boldsymbol{\beta}}^{or})\ge Q(\boldsymbol{\beta})-Q(\tilde{\boldsymbol{\beta}}^{or})+$$

$$\sum_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}, k\neq k'} \Big(\rho_\tau\left(\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_2,\lambda_2\right)-\rho_\tau(\|\tilde{\boldsymbol{\beta}}_i^{or}-\tilde{\boldsymbol{\beta}}_j^{or}\|_2,\lambda_2)\Big).$$

Since the minimum distance b satisfies $cb \geq \tau\lambda_2$, then we have

$$L(\boldsymbol{\beta}) - L(\tilde{\boldsymbol{\beta}}^{or}) \geq Q(\boldsymbol{\beta}) - Q(\tilde{\boldsymbol{\beta}}^{or}).$$

As $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathcal{M}^{\boldsymbol{\beta}}} Q(\boldsymbol{\beta})$, then $Q(\boldsymbol{\beta}) > Q(\tilde{\boldsymbol{\beta}}^{or})$ for any $\boldsymbol{\beta}$ such that $\big\|\mathbf{B}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}^{or})\big\|_n^2 = C(\frac{m}{n_0} + \frac{\lambda_1^2 m^{2d}}{n_0^2} + m^{-2d})$ if C is sufficiently large. This completes the proof. □

Proof of Corollary 3.1. For any trajectories i and j belonging to the same subgroup, we have $\|\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j\|_n^2 \leq \|\hat{\mathbf{f}}_i - \mathbf{f}_i^0\|_n^2 + \|\hat{\mathbf{f}}_j - \mathbf{f}_j^0\|_n^2 + \|\mathbf{f}_i^0 - \mathbf{f}_j^0\|_n^2 \leq 2\max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i\|_n^2 \leq O_p(\frac{m}{n_0}) + O_p(\frac{\lambda_1^2}{n_0^2} m^{2d}) + O_p(m^{-2q})$. As $n_0 \to \infty$, it is straightforward that $\|\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j\|_n^2 \leq O_p(m^{-2q})$. On the other hand, for any $i \in \mathcal{G}_k, j \in \mathcal{G}'_k, k \neq k'$, we have $\big\|\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j\big\|_n^2 \geq \min \|\mathbf{f}_i^0 - \mathbf{f}_j^0\|_n^2 - 2\max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^0\|_n^2 \geq b^2 - O_p(m^{-2q})$ as $n_0 \to \infty$. This completes the proof. □

Acknowledgements

This research was supported by National Science Foundation Grants (DMS-1308227, DMS1415308 and DMS1613190).

References

[1] Abraham, C., Cornillon, P.-A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics* **30**, 3, 581–595.

[2] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3**, 1, 1–122.

[3] Bronnenberg, B. J., Kruger, M. W., and Mela, C. F. (2008). Database paper - the iri marketing data set. *Marketing Science* **27**, 4, 745–748.

[4] Burren, O. S., Rubio García, A., Javierre, B.-M., Rainbow, D. B., Cairns, J., Cooper, N. J., Lambourne, J. J., Schofield, E., Castro Dopico, X., Ferreira, R. C., Coulson, R., Burden, F., Rowlston, S. P., Downes, K., Wingett, S. W., Frontini, M., Ouwehand, W. H., Fraser, P., Spivakov, M., Todd, J. A., Wicker, L. S., Cutler, A. J., and Wallace, C. (2017). Chromosome contacts in activated t cells identify autoimmune disease candidate genes. *Genome Biology* **18**, 1 (Sep), 165.

[5] Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics* **24**, 4, 994–1013.

[6] Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96**, 3, 529–544.

[7] Coffey, N., Hinde, J., and Holian, E. (2014). Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis 71*, 14–29.

[8] De Boor, C. (2001). *A practical guide to splines (revised ed.).* New York, Springer.

[9] Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science* **11**, 2, 89–102.

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 456, 1348–1360.

[11] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 458, 611–631.

[12] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**, 1, 100–108.

[13] Hastie, T., Tibshirani, R., and Walther, G. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society. Series B 63*, 411–423.

[14] Hsu, Y.-H., Zillikens, M. C., Wilson, S. G., Farber, C. R., Demissie, S., Soranzo, N., Bianchi, E. N., Grundberg, E., Liang, L., Richards, J. B., and others. (2010). An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits. *PLoS Genetics* **6**, 6.

[15] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2**, 1, 193–218.

[16] Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist* **11**, 2, 37–50.

[17] Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics* **19**, 4, 474–482.

[18] Ma, P., Castillo-Davis, C. I., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34**, 4, 1261–1269.

[19] Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* **112**, 517, 410–423.

[20] Pan, W., Shen, X., and Liu, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *The Journal of Machine Learning Research* **14**, 1, 1865–1889.

[21] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 336, 846–850.

[22] Ruppert, D. (2002). Selecting the number of knots for penalized

splines. *Journal of Computational and Graphical Statistics* **11**, 4, 735–757. MR1944261

[23] Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107**, 497, 223–232. MR2949354

[24] Wu, J., Zhu, J., Wang, L., and Wang, S. (2017). Genome-wide association study identifies nbs-lrr-encoding genes related with anthracnose and common bacterial blight in the common bean. *Frontiers in Plant Science 8*, 1398.

[25] Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks* **16**, 3, 645–678.

[26] Xue, L., Qu, A., and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association* **105**, 492, 1518–1530. MR2796568

[27] Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. *Statistica Sinica* **16**, 4, 1423–1446. MR2327498

[28] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 1, 121–130.

[29] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 2, 894–942. MR2604701

[30] Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics* **26**, 5, 1760–1782. MR1673277