Journal of Multivariate Analysis 189 (2022) 104874

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Biclustering analysis of functionals via penalized fusion

Kuangnan Fang [a], Yuanxing Chen [a], Shuangge Ma [b], Qingzhao Zhang [c,*]

[a] *Department of Statistics and Data Science, School of Economics, Xiamen University, China*
[b] *Department of Biostatistics, Yale University, United States of America*
[c] *MOE Key Laboratory of Econometrics, Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, China*

ARTICLE INFO

Article history:
Received 4 November 2020
Received in revised form 12 June 2021
Accepted 20 October 2021
Available online 29 October 2021

AMS 2020 subject classifications:
primary 62H30
62R10

Keywords:
Biclustering
Functional data
Penalized fusion

ABSTRACT

In biomedical data analysis, clustering is commonly conducted. Biclustering analysis conducts clustering in both the sample and covariate dimensions and can more comprehensively describe data heterogeneity. In most of the existing biclustering analyses, scalar measurements are considered. In this study, motivated by time-course gene expression data and other examples, we take the "natural next step" and consider the biclustering analysis of functionals under which, for each covariate of each sample, a function (to be exact, its values at discrete measurement points) is present. We develop a doubly penalized fusion approach, which includes a smoothness penalty for estimating functionals and, more importantly, a fusion penalty for clustering. Statistical properties are rigorously established, providing the proposed approach a strong ground. We also develop an effective ADMM algorithm and accompanying R code. Numerical analysis, including simulations, comparisons, and the analysis of two time-course gene expression data, demonstrates the practical effectiveness of the proposed approach.

1. Introduction

In biomedical data analysis, clustering has been routinely conducted. The clustering of samples can assist better understanding sample heterogeneity, and the clustering of covariates can identify those that behave similarly across samples and then, for example, improve our understanding of covariate functionalities. Clustering can also serve as the basis of other analysis, for example, regression. Biclustering analysis has also been developed, identifying clustering structures in both sample and covariate dimensions. It includes sample- and covariate-clustering as special cases and, in a sense, can be more comprehensive. For generic reviews of techniques, theories, and applications of clustering, we refer to [19,46].

This study has been partly motivated by the analysis of gene expression data, for which sample- and covariate-clustering as well as biclustering have been extensively conducted [21,45]. Most gene expression studies generate "snapshot" values. Unlike some types of omics measurements, gene expression values can be time-dependent, and the temporal trends of gene expressions can have important biological implications [16]. Accordingly, time-course gene expression studies have been conducted, generating multiple measurements at different time points for each gene of each sample. In the analysis of time-course gene expression data, besides simple statistics, functional data analysis (FDA) techniques, have been adopted and shown as powerful [12].

[*] Corresponding author.
E-mail address: zhangqingzhao@amss.ac.cn (Q. Zhang).

https://doi.org/10.1016/j.jmva.2021.104874

FDA deals with data samples that consist of curves or other infinite-dimensional data objects. Over the last two decades, we have witnessed significant developments in its theory, method, computation, and application. For systematic reviews, we refer to [2,15,23,40]. In FDA, clustering analysis has been of particular interest. A popular approach projects functional data into a finite-dimensional space and then applies existing clustering methods. For example, Abraham et al. [1] conduct B-spline expansions, and clusters the estimated coefficients using a k-means algorithm. Peng and Müller [30] develop a distance for sparse functional data, and apply a k-means algorithm to functional principle component analysis (PCA) scores. Other approaches, such as Bayesian [37], subspace [3,9,10], and model-based [18,20], have also been developed. We refer to [17,40] for surveys on functional data clustering. Most works in this area, however, have focused on either sample- or covariate-clustering.

For biclustering analysis (of gene expression and other types of data), in this article, we take the "natural next step" and consider the scenario where for each covariate of each sample, a function or its realizations at discrete time points are available. We note that, although this study has been partly motivated by gene expression data and some of the discussions are focused on such data, the considered data scenario and proposed technique can have applications far beyond such data. For example, in biomedical studies, many biomarkers measured in blood tests vary across time, and their values can be obtained from medical records. In financial studies, many measures of a company, for example size and stock price, vary across time. As such, our investigation can have broad applications.

There is a vast literature on biclustering analysis with scalar measurements. Directly applying such techniques to the present problem will involve either treating functional measurements as scalars and then computing distances (between covariates and samples) – which may be ineffective by not sufficiently accounting for the functional nature of data, or first estimating functionals and then computing distances between the estimates – which may also encounter challenges when a large number of functionals need to be jointly estimated. Our literature review suggests that there are also a handful recent biclustering methods designed for functional (especially including longitudinal) data. For example, Slimen et al. [35] propose a biclustering method for multivariate functional data based on the Gaussian latent block model (LBM) using the first functional PCA scores. Bouveyron et al. [4] develop an extension of the Gaussian LBM by modeling the whole set of functional PCA scores. In another work [28], a biclustering method with a plaid model is extended to three-dimensional data arrays, of which multivariate longitudinal data is a special case.

For the biclustering analysis of functionals, in this article, we develop a penalized fusion based approach. More specifically, a nonparametric model is assumed for each covariate of each sample, allowing for sufficient flexibility in modeling. A doubly penalization technique is adopted, which includes a smoothness penalty to regulate nonparametric estimation. The most significant advancement is the second, fusion penalty, which "transforms" clustering in both sample and covariate dimensions to a penalized estimation problem. Statistical and numerical investigations are conducted, providing the proposed approach a solid ground. This study may complement and advance from the existing ones in multiple aspects. Compared to direct applications of biclustering methods for scalars (that either directly compute distances without functional estimation or estimate functionals separately), the proposed approach can more effectively accommodate the functional nature of data or generate more effective estimation. This is because it "combines" clustering and estimation, and as such, estimation only needs to be conducted for clusters as opposed to individual covariates, potentially leading to a smaller number of parameters and hence more effective estimation. Compared to some of the existing biclustering methods for functionals, such as [4,35], the proposed approach has a much easier way of determining the number of clusters. In addition, unlike [4,35], it does not make stringent distributional assumptions (for example, normality). Meanwhile, rigorous theoretical investigations are conducted beyond methodological developments, granting the proposed approach a stronger statistical basis. It also advances from the clustering of functional covariate effects (assuming homogeneous samples) by simultaneously examining sample heterogeneity, thus being more comprehensive. Additionally, this study may also advance and enrich the penalized fusion technique. Clustering via penalized fusion has been pioneered in [8] and other studies. Compared to alternative clustering techniques, it is more recent and has notable statistical and numerical advantages [44]. Compared to the existing penalized fusion based clustering, this study differs by conducting biclustering and by having unknown parameters generated from the basis expansion of functionals. Last but not least, this study also provides a practically useful and new way of analyzing time-course gene expression data (and other data with similar characteristics).

The remainder of this article is organized as follows: Section 2 introduces the new biclustering approach via penalized fusion and develops an effective computational algorithm. Statistical properties are established to provide our method a strong theoretical support. Simulation studies and the analysis of two time-course expression data are conducted in Sections 3 and 4, respectively. Section 5 concludes with a brief discussion. The proofs of the main results are presented in Appendix A.

2. Methods

2.1. Data and model settings

For the $j \in \{1, \ldots, q\}$th covariate of sample $i \in \{1, \ldots, N\}$, denote $\mathbf{Y}_{i,j} = (Y_{i,j,1}, \ldots, Y_{i,j,n_{i,j}})^\top$ as the ordered measurements (ordered by time for time-course gene expression data), which are the discrete realizations of an unknown underlying functional. Further denote $\mathbf{Y}_i = (\mathbf{Y}_{i,1}^\top, \dots, \mathbf{Y}_{i,q}^\top)^\top$, $\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_N^\top)^\top$, and $n = \sum_{i=1}^N \sum_{j=1}^q n_{i,j}$. Under the biclustering analysis framework, assume that data can be "decomposed" into K_r sample (row) groups and K_c covariate (column) groups. Note that advancing from many existing approaches, the numbers of two dimensional groups are not pre-specified. Denote $t_{i,j,m}'s \in \mathcal{T} = [0, 1]$ as the observed time points. If (sample i, covariate j) belongs to the k_rth sample group and the k_cth covariate group, then

$$Y_{i,j,m} = g_{(k_r,k_c)}(t_{i,j,m}) + \epsilon_{i,j,m}, \tag{1}$$

where $g_{(k_r,k_c)}(t)$ is the unknown mean function, and $\epsilon_{i,j,m}'s$ are the random errors with mean zero.

For estimation, we adopt the basis expansion technique. Specifically, denote $U_p(t) = (U_{1,p}(t), \dots, U_{p,p}(t))^\top$ as the collection of p rescaled basis functions. In the literature, there are extensive studies on choosing the form and number of basis functions [32], which will not be reiterated here. In our numerical study, we adopt B-spline basis functions of order $d = 3$. Let $g_{i,j}(t)$ be the unknown mean function for the jth covariate of the ith sample, then we have

$$g_{i,j}(t) \approx U_p^\top(t)\boldsymbol{\beta}_{i,j},$$

where $\boldsymbol{\beta}_{i,j} = (\beta_{i,j,1}, \dots, \beta_{i,j,p})^\top$ is the vector of unknown coefficients. Further denote $\mathbf{U}_{i,j} = (U_p(t_{i,j,1}), \dots, U_p(t_{i,j,n_{i,j}}))^\top$. For estimation (without clustering), consider the objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\boldsymbol{\beta}\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^\top\mathbf{M}\boldsymbol{\beta} = \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^q\Big(\|\mathbf{Y}_{i,j} - \mathbf{U}_{i,j}\boldsymbol{\beta}_{i,j}\|_2^2 + \gamma_1\boldsymbol{\beta}_{i,j}^\top\mathbf{D}\boldsymbol{\beta}_{i,j}\Big), \tag{2}$$

where $\mathbf{U} = \mathrm{diag}(\mathbf{U}_{1,1}, \dots, \mathbf{U}_{1,q}, \dots, \mathbf{U}_{N,q})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,1}^\top, \dots, \boldsymbol{\beta}_{1,q}^\top, \dots, \boldsymbol{\beta}_{N,q}^\top)^\top$, $\mathbf{M} = \mathrm{diag}(\mathbf{D}, \dots, \mathbf{D})$, $\mathbf{D} = \boldsymbol{\delta}^\top\boldsymbol{\delta}$, $\boldsymbol{\delta}$ is a $(p-2)\times p$ matrix representing the second order differential operator, and γ_1 is a non-negative tuning parameter. In this objective function, the first term is the lack-of-fit, and the penalty term controls the smoothness of estimation.

2.2. Biclustering via penalized fusion

Under the clustering via penalized fusion framework, two samples (covariates) belong to the same cluster if and only if they have the same regression coefficients. As such, clustering amounts to determining whether two samples (covariates) have the same estimated coefficients. For samples $i_1, i_2 \in \{1, \dots, N\}$, denote $\boldsymbol{\beta}_{i_1}^{(r)}, \boldsymbol{\beta}_{i_2}^{(r)}$ as the length $p \times q$ vectors of coefficients. For covariates $j_1, j_2 \in \{1, \dots, q\}$, denote $\boldsymbol{\beta}_{j_1}^{(c)}, \boldsymbol{\beta}_{j_2}^{(c)}$ as the length $p \times N$ vectors of coefficients. For estimating $\boldsymbol{\beta}$ and hence determining the clustering structure, we propose minimizing the objective function:

$$L(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + \sum_{1\le i_1<i_2\le N} p_\tau(\|\boldsymbol{\beta}_{i_1}^{(r)} - \boldsymbol{\beta}_{i_2}^{(r)}\|_2, \gamma_2) + \sum_{1\le j_1<j_2\le q} p_\tau(\|\boldsymbol{\beta}_{j_1}^{(c)} - \boldsymbol{\beta}_{j_2}^{(c)}\|_2, (N/q)^{1/2}\gamma_2). \tag{3}$$

Here $p_\tau(,)$ is a penalty function, τ is a regularization parameter, $\|\cdot\|_2$ is the ℓ_2 norm, and γ_2 is a data-dependent tuning parameter. $(N/q)^{1/2}$ is added to make the two penalties comparable. In our numerical study, we adopt MCP [47], that is, $p_\tau(t,\gamma) = \gamma\int_0^t(1 - x/(\tau\gamma))_+ dx$ with $\tau > 1$. Here $(x)_+ = x$ if $x > 0$, and $(x)_+ = 0$ otherwise. Note that SCAD [14] and some other penalties are also applicable. Denote the estimator as $\hat{\boldsymbol{\beta}}$. Let $\{\hat{\boldsymbol{\alpha}}_1^{(r)}, \dots, \hat{\boldsymbol{\alpha}}_{\hat{K}_r}^{(r)}\}$ be the distinct values of $\hat{\boldsymbol{\beta}}_i^{(r)}$'s. Similarly, let $\{\hat{\boldsymbol{\alpha}}_1^{(c)}, \dots, \hat{\boldsymbol{\alpha}}_{\hat{K}_c}^{(c)}\}$ be the distinct values of $\hat{\boldsymbol{\beta}}_j^{(c)}$'s. We can then obtain the block structure of $\hat{\boldsymbol{\beta}}$ by $\{\hat{\boldsymbol{\alpha}}_{1,1}^{(r,c)}, \dots, \hat{\boldsymbol{\alpha}}_{\hat{K}_r,\hat{K}_c}^{(r,c)}\}$, which are the distinct values of $\hat{\boldsymbol{\beta}}_{i,j}$, and set $\hat{K}_b = \hat{K}_r \times \hat{K}_c$.

In (3), penalty is imposed to the norms of all pairwise differences to promote equality, as in "standard" penalized fusion [8]. Here it is noted that, as in [8], since there is no information on the order of samples/covariates, all pairwise differences are taken, which differs from, for example, fused Lasso and other fused penalizations. Different from [8], as clustering needs to be conducted in both the sample and covariate dimensions, two fusion penalties are imposed, promoting equality in two directions. It is also noted that each specific coefficient shows up in three different penalties. As to be shown below, with properly chosen tunings, there is not an over penalization problem. In addition, it is not rare to have a parameter involved in two or more penalties [7].

The proposed approach involves two tunings, which have "ordinary" implications, with one controlling smoothness and the other determining the structure of clustering. One possibility is to conduct a two-dimensional grid search. Here we adopt the alternative proposed in [48], which has two steps and a lower computational cost. In particular, in the first step, we set $\gamma_2 = 0$ and select the optimal γ_1 by minimizing:

$$\mathrm{BIC}(\gamma_1) = \sum_{i=1}^N\sum_{j=1}^q\left\{\log\left(\frac{\|\mathbf{Y}_{i,j} - \hat{\mathbf{g}}_{i,j}\|_2^2}{n_{i,j}}\right) + \frac{\log(n_{i,j})}{n_{i,j}}\mathrm{df}_{i,j}\right\},$$

where $\mathrm{df}_{i,j} = \mathrm{trace}\Big\{\mathbf{U}_{i,j}(\mathbf{U}_{i,j}^\top\mathbf{U}_{i,j} + \gamma_1\mathbf{D})^{-1}\mathbf{U}_{i,j}^\top\Big\}$ and $\hat{\mathbf{g}}_{i,j} = (\hat{g}_{i,j}(t_{i,j,1}), \dots, \hat{g}_{i,j}(t_{i,j,n_{i,j}}))^\top$ with $\hat{g}_{i,j}(t) = U_p^\top(t)\hat{\boldsymbol{\beta}}_{i,j}$.

In the second step, we fix the value of γ_1 at the optimal and select γ_2 by minimizing

$$\mathrm{BIC}(\gamma_2) = \log\left(\frac{\|\mathbf{Y} - \hat{\boldsymbol{g}}\|_2^2}{Nq}\right) + \frac{\log(Nq)}{Nq}\mathrm{df},$$

where $\mathrm{df} = (\hat{K}_r\hat{K}_c/Nq)\sum_{i=1}^{N}\sum_{j=1}^{q}\mathrm{df}_{i,j}$ and $\hat{\boldsymbol{g}} = (\hat{\boldsymbol{g}}_{1,1}^\top, \ldots, \hat{\boldsymbol{g}}_{N,q}^\top)^\top$.

2.3. Computation

We develop an effective algorithm based on the ADMM technique. Specifically, we first reformulate (3) as

$$\begin{aligned} &\arg\min \quad Q(\boldsymbol{\beta}) + \sum_{\delta\in\Delta^{(r)}} p_\tau(\|\boldsymbol{\eta}_\delta^{(r)}\|_2, \gamma_2) + \sum_{\delta\in\Delta^{(c)}} p_\tau(\|\boldsymbol{\eta}_\delta^{(c)}\|_2, (N/q)^{1/2}\gamma_2),\\ &\text{subject to} \quad \boldsymbol{\beta}_{i_1}^{(r)} - \boldsymbol{\beta}_{i_2}^{(r)} = \boldsymbol{\eta}_\delta^{(r)}, \qquad \boldsymbol{\beta}_{j_1}^{(c)} - \boldsymbol{\beta}_{j_2}^{(c)} = \boldsymbol{\eta}_\delta^{(c)}, \end{aligned}$$

where $\Delta^{(r)} = \{\delta = (i_1, i_2) : 1 \le i_1 < i_2 \le N\}$ and $\Delta^{(c)} = \{\delta = (j_1, j_2) : 1 \le j_1 < j_2 \le q\}$. Optimizing the constrained objective function is equivalent to optimizing the augmented Lagrangian function:

$$\begin{aligned} L_\theta(\boldsymbol{\beta}, \mathbf{H}_r, \mathbf{H}_c, \boldsymbol{\Lambda}_r, \boldsymbol{\Lambda}_c) = {} & \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\boldsymbol{\beta}\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^\top\mathbf{M}\boldsymbol{\beta} + \sum_{\delta\in\Delta^{(r)}} p_\tau(\|\boldsymbol{\eta}_\delta^{(r)}\|_2, \gamma_2) + \sum_{\delta\in\Delta^{(r)}} \boldsymbol{\lambda}_\delta^{(r)\top}(\boldsymbol{\eta}_\delta^{(r)} - \boldsymbol{\beta}_{i_1}^{(r)} + \boldsymbol{\beta}_{i_2}^{(r)})\\ & + \frac{\theta}{2}\sum_{\delta\in\Delta^{(r)}} \|\boldsymbol{\eta}_\delta^{(r)} - \boldsymbol{\beta}_{i_1}^{(r)} + \boldsymbol{\beta}_{i_2}^{(r)}\|_2^2 + \sum_{\delta\in\Delta^{(c)}} p_\tau(\|\boldsymbol{\eta}_\delta^{(c)}\|_2, (N/q)^{1/2}\gamma_2) \qquad (4)\\ & + \sum_{\delta\in\Delta^{(c)}} \boldsymbol{\lambda}_\delta^{(c)\top}(\boldsymbol{\eta}_\delta^{(c)} - \boldsymbol{\beta}_{j_1}^{(c)} + \boldsymbol{\beta}_{j_2}^{(c)}) + \frac{\theta}{2}\sum_{\delta\in\Delta^{(c)}} \|\boldsymbol{\eta}_\delta^{(c)} - \boldsymbol{\beta}_{j_1}^{(c)} + \boldsymbol{\beta}_{j_2}^{(c)}\|_2^2, \end{aligned}$$

where θ is a small positive constant, $\mathbf{H}_r = (\boldsymbol{\eta}_{(1,2)}^{(r)}, \ldots, \boldsymbol{\eta}_{(N-1,N)}^{(r)})$, $\mathbf{H}_c = (\boldsymbol{\eta}_{(1,2)}^{(c)}, \ldots, \boldsymbol{\eta}_{(q-1,q)}^{(c)})$, $\boldsymbol{\Lambda}_r = (\boldsymbol{\lambda}_{(1,2)}^{(r)}, \ldots, \boldsymbol{\lambda}_{(N-1,N)}^{(r)})$, and $\boldsymbol{\Lambda}_c = (\boldsymbol{\lambda}_{(1,2)}^{(c)}, \ldots, \boldsymbol{\lambda}_{(q-1,q)}^{(c)})$. Here we introduce the dual variables $\boldsymbol{\lambda}_\delta^{(r)}$ and $\boldsymbol{\lambda}_\delta^{(c)}$ corresponding to the pair δ in $\Delta^{(r)}$ and $\Delta^{(c)}$, and the cardinality of $\Delta^{(r)}$ and $\Delta^{(c)}$ are denoted by $|\Delta^{(r)}|$ and $|\Delta^{(c)}|$.

We consider an iterative algorithm, where the updates in step $m+1$ are:

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} &= \arg\min_{\boldsymbol{\beta}} L_\theta\left(\boldsymbol{\beta}, \mathbf{H}_r^{(m)}, \mathbf{H}_c^{(m)}, \boldsymbol{\Lambda}_r^{(m)}, \boldsymbol{\Lambda}_c^{(m)}\right), \quad \mathbf{H}_r^{(m+1)} = \arg\min_{\mathbf{H}_r} L_\theta\left(\boldsymbol{\beta}^{(m+1)}, \mathbf{H}_r, \boldsymbol{\Lambda}_r^{(m)}\right),\\ \mathbf{H}_c^{(m+1)} &= \arg\min_{\mathbf{H}_c} L_\theta\left(\boldsymbol{\beta}^{(m+1)}, \mathbf{H}_c, \boldsymbol{\Lambda}_c^{(m)}\right), \quad \boldsymbol{\lambda}_\delta^{(r)(m+1)} = \boldsymbol{\lambda}_\delta^{(r)(m)} + \theta\left(\boldsymbol{\eta}_\delta^{(r)(m+1)} - \boldsymbol{\beta}_{i_1}^{(r)(m+1)} + \boldsymbol{\beta}_{i_2}^{(r)(m+1)}\right), \ \delta \in \Delta^{(r)}, \qquad (5)\\ \boldsymbol{\lambda}_\delta^{(c)(m+1)} &= \boldsymbol{\lambda}_\delta^{(c)(m)} + \theta\left(\boldsymbol{\eta}_\delta^{(c)(m+1)} - \boldsymbol{\beta}_{j_1}^{(c)(m+1)} + \boldsymbol{\beta}_{j_2}^{(c)(m+1)}\right), \ \delta \in \Delta^{(c)}. \end{aligned}$$

More specifically, when optimizing over $\boldsymbol{\beta}$, we consider

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\boldsymbol{\beta}\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^\top\mathbf{M}\boldsymbol{\beta} + \frac{\theta}{2}\Big(\sum_{\delta\in\Delta^{(r)}} \|\tilde{\boldsymbol{\eta}}_\delta^{(r)(m)} - \mathbf{B}_\delta^{(r)}\boldsymbol{\beta}\|_2^2 + \sum_{\delta\in\Delta^{(c)}} \|\tilde{\boldsymbol{\eta}}_\delta^{(c)(m)} - \mathbf{B}_\delta^{(c)}\boldsymbol{\beta}\|_2^2\Big), \qquad (6)$$

where $\tilde{\boldsymbol{\eta}}_\delta^{(r)} = \boldsymbol{\eta}_\delta^{(r)} + \frac{1}{\theta}\boldsymbol{\lambda}_\delta^{(r)}$, $\tilde{\boldsymbol{\eta}}_\delta^{(c)} = \boldsymbol{\eta}_\delta^{(c)} + \frac{1}{\theta}\boldsymbol{\lambda}_\delta^{(c)}$, $\mathbf{B}_\delta^{(r)} = (\mathbf{e}_{i_1}^{(r)} - \mathbf{e}_{i_2}^{(r)})^\top \otimes \mathbf{I}_{qp}$, $\mathbf{B}_\delta^{(c)} = \mathbf{I}_N \otimes [(\mathbf{e}_{j_1}^{(c)} - \mathbf{e}_{j_2}^{(c)})^\top \otimes \mathbf{I}_p]$, $\mathbf{e}_i^{(r)}$ is an $N \times 1$ zero vector except that its ith element is 1, $\mathbf{e}_j^{(c)}$ is a $q \times 1$ zero vector except that its jth element is 1, $\otimes$ is the Kronecker product, and $\mathbf{I}_p$ is a $p \times p$ identity matrix. Denote $\mathbf{B}_r = (\mathbf{B}_{(1,2)}^{(r)\top}, \ldots, \mathbf{B}_{(N-1,N)}^{(r)\top})^\top$, $\mathbf{B}_c = (\mathbf{B}_{(1,2)}^{(c)\top}, \ldots, \mathbf{B}_{(q-1,q)}^{(c)\top})^\top$, $\tilde{\mathbf{H}}_r = (\tilde{\boldsymbol{\eta}}_{(1,2)}^{(r)}, \ldots, \tilde{\boldsymbol{\eta}}_{(N-1,N)}^{(r)})$, and $\tilde{\mathbf{H}}_c = (\tilde{\boldsymbol{\eta}}_{(1,2)}^{(c)}, \ldots, \tilde{\boldsymbol{\eta}}_{(q-1,q)}^{(c)})$. Then the update for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{U}^\top\mathbf{U} + \gamma_1\mathbf{M} + \theta\mathbf{B}_r^\top\mathbf{B}_r + \theta\mathbf{B}_c^\top\mathbf{B}_c\right)^{-1}\left(\mathbf{U}^\top\mathbf{Y} + \theta\mathbf{B}_r^\top\mathrm{vec}\left(\tilde{\mathbf{H}}_r^{(m)}\right) + \theta\mathbf{B}_c^\top\mathrm{vec}\left(\tilde{\mathbf{H}}_c^{(m)}\right)\right), \qquad (7)$$

where $\mathrm{vec}(\mathbf{Z})$ is the vectorization of matrix $\mathbf{Z}$ by columns.

For $\mathbf{H}_r$, we consider

$$f(\boldsymbol{\eta}_\delta^{(r)}) = p_\tau(\|\boldsymbol{\eta}_\delta^{(r)}\|_2, \gamma_2) + \frac{\theta}{2}\left\|\boldsymbol{\eta}_\delta^{(r)} - \boldsymbol{\beta}_{i_1}^{(r)(m+1)} + \boldsymbol{\beta}_{i_2}^{(r)(m+1)} + \boldsymbol{\lambda}_\delta^{(r)(m)}/\theta\right\|_2^2. \qquad (8)$$

Denote $\boldsymbol{z}_\delta^{(r)(m+1)} = \boldsymbol{\beta}_{i_1}^{(r)(m+1)} - \boldsymbol{\beta}_{i_2}^{(r)(m+1)} - \boldsymbol{\lambda}_\delta^{(r)(m)}/\theta$. With the KKT conditions of (8), we can get a closed form solution of $\mathbf{H}_r$:

$$\boldsymbol{\eta}_\delta^{(r)(m+1)} = \begin{cases} \boldsymbol{z}_\delta^{(r)(m+1)}, & \text{if } \|\boldsymbol{z}_\delta^{(r)(m+1)}\|_2 \ge \tau\gamma_2,\\ \dfrac{\tau\theta}{\tau\theta - 1}\left(1 - \dfrac{\gamma_2/\theta}{\|\boldsymbol{z}_\delta^{(r)(m+1)}\|_2}\right)_+\boldsymbol{z}_\delta^{(r)(m+1)}, & \text{if } \|\boldsymbol{z}_\delta^{(r)(m+1)}\|_2 < \tau\gamma_2. \end{cases} \qquad (9)$$

Similarly, denote $\boldsymbol{z}_{\delta}^{(c)(m+1)} = \boldsymbol{\beta}_{j_1}^{(c)(m+1)} - \boldsymbol{\beta}_{j_2}^{(c)(m+1)} - \boldsymbol{\lambda}_{\delta}^{(c)(m)}/\theta$, and we can get a closed form solution of $\mathbf{H}_c$:

$$\boldsymbol{\eta}_{\delta}^{(c)(m+1)} = \begin{cases} \boldsymbol{z}_{\delta}^{(c)(m+1)}, & \text{if } \|\boldsymbol{z}_{\delta}^{(c)(m+1)}\|_2 \geq (N/q)^{1/2}\tau\gamma_2, \\ \dfrac{\tau\theta}{\tau\theta-1}\left(1-\dfrac{(N/q)^{1/2}\gamma_2/\theta}{\|\boldsymbol{z}_{\delta}^{(c)(m+1)}\|_2}\right)_+ \boldsymbol{z}_{\delta}^{(c)(m+1)}, & \text{if } \|\boldsymbol{z}_{\delta}^{(c)(m+1)}\|_2 < (N/q)^{1/2}\tau\gamma_2. \end{cases} \tag{10}$$

Consider the initial values $\boldsymbol{\beta}^{(0)} = (\mathbf{U}^{\top}\mathbf{U} + \gamma_1\mathbf{M})^{-1}\mathbf{U}^{\top}\mathbf{Y}$, $\boldsymbol{\eta}_{\delta}^{(r)(0)} = \boldsymbol{\beta}_{i_1}^{(r)(0)} - \boldsymbol{\beta}_{i_2}^{(r)(0)}$, and $\boldsymbol{\eta}_{\delta}^{(c)(0)} = \boldsymbol{\beta}_{j_1}^{(c)(0)} - \boldsymbol{\beta}_{j_2}^{(c)(0)}$, and $\boldsymbol{\Lambda}_r^{(0)}$ and $\boldsymbol{\Lambda}_c^{(0)}$ are set as zero. The ADMM based algorithm is summarized in Algorithm 1.

Algorithm 1

```
Input:
    Response vector $\mathbf{Y}$, basis expansion design matrix $\mathbf{U}$, and difference matrix $\mathbf{M}$;
    Tuning parameters $\gamma_1$ and $\gamma_2$. Specific to MCP, regularization parameter $\tau$;
Output:
    Coefficient vector $\boldsymbol{\beta}$, splitting variables $\mathbf{H}_r$ and $\mathbf{H}_c$, and dual variables $\boldsymbol{\Lambda}_r$ and $\boldsymbol{\Lambda}_c$;
repeat
    for $m = 0, 1, 2 \cdots$ do
        Update $\boldsymbol{\beta}$ by (7).
        Update $\mathbf{H}_r$ by (9).
        Update $\mathbf{H}_c$ by (10).
        Update $\boldsymbol{\Lambda}_r$ and $\boldsymbol{\Lambda}_c$ by (5).
    end for
until the stopping criteria are met, which are set as $||\mathbf{r}_r^{(m+1)}||_2 \leq \epsilon_1^{pri}$, $||\mathbf{r}_c^{(m+1)}||_2 \leq \epsilon_2^{pri}$, $||\mathbf{s}_r^{(m+1)}||_2 \leq \epsilon_1^{dual}$, and $||\mathbf{s}_c^{(m+1)}||_2 \leq \epsilon_2^{dual}$ in our numerical study.
```

Proposition 1. *Denote the two primal residuals as* $\mathbf{r}_r^{(m+1)} = \mathbf{B}_r\boldsymbol{\beta}^{(m+1)} - vec(\mathbf{H}_r^{(m+1)})$ *and* $\mathbf{r}_c^{(m+1)} = \mathbf{B}_c\boldsymbol{\beta}^{(m+1)} - vec(\mathbf{H}_c^{(m+1)})$, *and the two dual residuals as* $\mathbf{s}_r^{(m+1)} = \theta\mathbf{B}_r^{\top}\left[vec(\mathbf{H}_r^{(m+1)}) - vec(\mathbf{H}_r^{(m)})\right]$ *and* $\mathbf{s}_c^{(m+1)} = \theta\mathbf{B}_c^{\top}\left[vec(\mathbf{H}_c^{(m+1)}) - vec(\mathbf{H}_c^{(m)})\right]$. *Then*

$$\lim_{m\to\infty} \|\mathbf{r}_r^{(m+1)}\|_2^2 = 0, \quad \lim_{m\to\infty} \|\mathbf{r}_c^{(m+1)}\|_2^2 = 0, \quad \lim_{m\to\infty} \|\mathbf{s}_r^{(m+1)} + \mathbf{s}_c^{(m+1)}\|_2^2 = 0.$$

This result establishes convergence of the proposed algorithm. In numerical analysis, we stop the algorithm and conclude convergence when $\|\mathbf{r}_r^{(m+1)}\|_2 \leq \epsilon_1^{pri}$, $\|\mathbf{r}_c^{(m+1)}\|_2 \leq \epsilon_2^{pri}$, $\|\mathbf{s}_r^{(m+1)}\|_2 \leq \epsilon_1^{dual}$ and $\|\mathbf{s}_c^{(m+1)}\|_2 \leq \epsilon_2^{dual}$. Following [5], we set the tolerance parameters as follows:

$$\begin{aligned} \epsilon_1^{pri} &= \sqrt{|\Delta^{(r)}|pq}\,\epsilon^{abs} + \epsilon^{rel}\max\left\{\|\mathbf{B}_r\boldsymbol{\beta}^{(m+1)}\|_2, \|\mathrm{vec}(\mathbf{H}_r^{(m+1)})\|_2\right\}, \\ \epsilon_2^{pri} &= \sqrt{|\Delta^{(c)}|pN}\,\epsilon^{abs} + \epsilon^{rel}\max\left\{\|\mathbf{B}_c\boldsymbol{\beta}^{(m+1)}\|_2, \|\mathrm{vec}(\mathbf{H}_c^{(m+1)})\|_2\right\}, \\ \epsilon_1^{dual} &= \sqrt{Nqp}\,\epsilon^{abs} + \epsilon^{rel}\|\mathbf{B}_r^{\top}\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m+1)})\|_2, \quad \epsilon_2^{dual} = \sqrt{Nqp}\,\epsilon^{abs} + \epsilon^{rel}\|\mathbf{B}_c^{\top}\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m+1)})\|_2. \end{aligned} \tag{11}$$

Here ϵ^{abs} and ϵ^{rel} are predetermined small values, for example 10^{-3}. In all of our numerical analysis, convergence is satisfactorily achieved within a small to moderate number of iterations. The code and example are publicly available at https://github.com/ruiqwy/Biclustering.

2.4. Statistical properties

For a vector $\boldsymbol{z} = (z_1, \ldots, z_s)^{\top} \in \mathbb{R}^s$, let $\|\boldsymbol{z}\|_{\infty} = \max_{1\leq l\leq s}|z_l|$. For a matrix $\mathbf{Z}_{s\times h}$, let $\|\mathbf{Z}\|_2 = \max_{\boldsymbol{v}\in\mathbb{R}^h, \|\boldsymbol{v}\|_2=1}\|\mathbf{Z}\boldsymbol{v}\|_2$ and $\|\mathbf{Z}\|_{\infty} = \max_{1\leq i\leq s}\sum_{j=1}^{h}|Z_{i,j}|$. For any two sequences of real numbers $\{a_n\} \geq 1$ and $\{b_n\} \geq 1$, denote $b_n \ll a_n$ if $b_n/a_n = o(1)$. Let r be a positive integer, $v \in (0, 1]$, and $\kappa = r + v > 1.5$. Let $\mathcal{H}$ be the collection of functions g on $\mathcal{T} = [0, 1]$, where the rth derivative $g^{(r)}$ exists and satisfies the Lipschitz condition with order v:

$$|g^{(r)}(z_1) - g^{(r)}(z_2)| \leq C|z_1 - z_2|^{v}, \quad 0 \leq z_1, z_2 \leq 1,$$

and C is a positive constant.

Define the following collections of index sets for clustering memberships: $\mathcal{G}^{(r)} = (\mathcal{G}_1^{(r)}, \ldots, \mathcal{G}_{K_r}^{(r)})$ for samples, $\mathcal{G}^{(c)} = (\mathcal{G}_1^{(c)}, \ldots, \mathcal{G}_{K_c}^{(c)})$ for covariates, and $\mathcal{G}^{(r,c)} = (\mathcal{G}_{1,1}^{(r,c)}, \ldots, \mathcal{G}_{k_r,k_c}^{(r,c)}, \ldots, \mathcal{G}_{K_r,K_c}^{(r,c)})$ for both samples and covariates. Define $\mathcal{M}_G = \{\boldsymbol{\beta} \in \mathbb{R}^{Nqp} : \boldsymbol{\beta}_{i_1,j_1} = \boldsymbol{\beta}_{i_2,j_2}$, for any $(i_1,j_1),(i_2,j_2) \in \mathcal{G}_{k_r,k_c}^{(r,c)}, 1 \leq k_r \leq K_r, 1 \leq k_c \leq K_c\}$. Let $|\mathcal{G}_{k_r}^{(r)}|$, $|\mathcal{G}_{k_c}^{(c)}|$, and $|\mathcal{G}_{k_r,k_c}^{(r,c)}|$ be the sizes of $\mathcal{G}_{k_r}^{(r)}$, $\mathcal{G}_{k_c}^{(c)}$, and $\mathcal{G}_{k_r,k_c}^{(r,c)}$, respectively. Further define $|\mathcal{G}_{min}^{(r)}| = \min_{1\leq k_r\leq K_r}|\mathcal{G}_{k_r}^{(r)}|$, $|\mathcal{G}_{min}^{(c)}| = \min_{1\leq k_c\leq K_c}|\mathcal{G}_{k_c}^{(c)}|$, and $|\mathcal{G}_{min}^{(r,c)}| = |\mathcal{G}_{min}^{(r)}| \times |\mathcal{G}_{min}^{(c)}|$. $|\mathcal{G}_{max}^{(r,c)}|$ can be defined accordingly. Let $\rho(t) = \gamma^{-1}p_{\tau}(t,\gamma)$. Assume the following conditions.

(C1) $g_{k_r,k_c} \in \mathcal{H}$ for all $k_r \in \{1,\ldots,K_r\}$, $k_c \in \{1,\ldots,K_c\}$, and $|\mathcal{G}^{(r,c)}_{max}|^{1/(2\kappa)} \ll p \ll |\mathcal{G}^{(r,c)}_{min}|^{1/3}$.

(C2) The distribution of $t_{i,j,m}$'s, $i \in \{1,\ldots,N\}, j \in \{1,\ldots,q\}, m \in \{1,\ldots,n_{i,j}\}$ follows a density function f_T, which is absolutely continuous. There exist constants c_1 and C_1 such that $0 < c_1 \le \min_{t\in\mathcal{T}} f_T(t) \le \max_{t\in\mathcal{T}} f_T(t) \le C_1 < \infty$.

(C3) $n_{i,j}$'s are uniformly bounded for all $i \in \{1,\ldots,N\}, j \in \{1,\ldots,q\}$.

(C4) $p_\tau(t,\gamma)$ is symmetric, non-decreasing, and concave in t for $t \in [0,\infty]$. There exists a constant $0 < a < \infty$ such that $\rho(t)$ is a constant for all $t \ge a\gamma$, and $\rho(0) = 0$. $\rho'(t)$ exists and is continuous except for a finite number of t and $\rho'(0+) = 1$.

(C5) Let $\boldsymbol{\epsilon}_{i,j} = (\epsilon_{i,j,1},\ldots,\epsilon_{i,j,n_{i,j}})^\top$, where $\epsilon_{i,j,m}$'s are independent across (i,j) (among different individual observational vectors) and correlated across m (within the same (i,j)). Furthermore, there exist $F > 0$ and $c_2 > 0$, such that for all $i \in \{1,\ldots,N\}$ and $j \in \{1,\ldots,q\}$,

$$\mathrm{E}\big(\exp\{F|n_{i,j}^{-1}\boldsymbol{\epsilon}_{i,j}^\top\boldsymbol{\epsilon}_{i,j}|^{1/2}\}\big) \le c_2.$$

Similar conditions have been assumed in the literature. The first condition in (C1) ensures that the Hölder's condition is satisfied [36]. The second condition in (C1) pertains to the growth rate of the number of internal knots, in a way similar to [25] and [24]. Condition (C2) assumes the boundedness of the density function, similarly to [48] and others. Conditions similar to (C3) have been commonly made. In the analysis of high-dimensional data, conditions similar to (C4) have been common, and it is easy to verify that MCP and SCAD satisfy (C4). Condition (C5) gives the boundedness condition for the error terms, and a similar condition can be found in [11].

When the true clustering structure is known, the oracle estimator for $\boldsymbol\beta$ can be defined as

$$\hat{\boldsymbol\beta}^{or} = \arg\min_{\boldsymbol\beta\in\mathcal{M}_G} \frac{1}{2} \sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c}} \Big\{\|\mathbf{Y}_{i,j}-\mathbf{U}_{i,j}\boldsymbol\beta_{i,j}\|_2^2+\gamma_1\boldsymbol\beta_{i,j}^\top\mathbf{D}\boldsymbol\beta_{i,j}\Big\},$$

where $\hat{g}^{or}_{(k_r,k_c)}$ is defined as the oracle estimator of $g_{(k_r,k_c)}$ based on $\hat{\boldsymbol\beta}^{or}$. Let $\boldsymbol\beta^*$ be the underlying true coefficient vector and $g^*_{(k_r,k_c)}$ be the true value of $g_{(k_r,k_c)}$. For any L^2-integrable function g, denote $\|g\| = (\int_{t\in\mathcal{T}} g^2(t)f_T(t)dt)^{1/2}$.

Theorem 1. *Assume that (C1)–(C5) hold. If* $\gamma_1 = o(|\mathcal{G}^{(r,c)}_{min}|^{-1/2})$ *and* $p\log(Nq) \ll |\mathcal{G}^{(r,c)}_{min}|$, *then with probability at least* $1 - 3K_rK_cp/(Nq)$,

$$\sup_{1\le i\le N, 1\le j\le q} \|\hat{\boldsymbol\beta}^{or}_{i,j} - \boldsymbol\beta^*_{i,j}\|_2 \le \psi, \qquad \sup_{1\le k_r\le K_r, 1\le k_c\le K_c} \|\hat{g}^{or}_{(k_r,k_c)} - g^*_{(k_r,k_c)}\| \le \psi,$$

where $\psi = C^*\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}$, *and* C^* *is a large constant.*

This theorem establishes consistency of the oracle estimates with a high probability. Denote $b = \min_{(k_r,k_c)\ne(k'_r,k'_c)} \|g^*_{(k_r,k_c)} - g^*_{(k'_r,k'_c)}\|$. We can further establish the following result.

Theorem 2. *Assume that (C1)–(C5) and conditions in Theorem 1 hold. If* $b \gg \gamma_2|\mathcal{G}^{(c)}_{min}|^{-1/2}$, $b \gg (N/q)^{1/2}\gamma_2|\mathcal{G}^{(r)}_{min}|^{-1/2}$, *and* $\gamma_2 \gg (pq)^{1/2}\log(Nq)/\min\{|\mathcal{G}^{(r)}_{min}|, |\mathcal{G}^{(c)}_{min}|\}$, *then there exists a local minimizer* $\hat{\boldsymbol\beta}$ *of* $L(\boldsymbol\beta)$ *satisfying*

$$P(\hat{\boldsymbol\beta} = \hat{\boldsymbol\beta}^{or}) \to 1 \ \text{ as } N, q \to \infty.$$

This theorem establishes that the oracle estimator is a local minimizer of the objective function with a high probability. The estimation consistency along with the separateness of the true functions can lead to the clustering consistency.

3. Simulation

We conduct simulation to assess performance of the proposed approach and gauge against the following alternatives: (a) the bKmeans method [1], which first fits each curve using B-splines and then clusters the estimated coefficients using the k-means technique by rows and columns, (b) the funHDDC method [33], which has been developed for multivariate functional data clustering based on latent mixture models. It has been applied to longitudinal data, and (c) the funLBM method [4], which has been developed for functional data biclustering based on latent block models. Here we note that the proposed and funLBM methods conduct biclustering directly, whereas the bKmeans and funHDDC methods have been originally designed for one-way clustering–hence they are applied twice to achieve both row and column clusterings. In addition, the funHDDC and funLBM methods are not directly applicable to functional data with unequal measurements. We apply imputation [26] to tackle this problem. As discussed in Section 1, biclustering methods for functional data are very limited. It is possible to modify other existing one-way functional clustering methods to achieve biclustering, however, this demands additional methodological developments. The three alternatives considered here have been chosen because of their closely related frameworks and competitive performance.

In evaluation, we examine both clustering and estimation accuracy. Specifically, when examining clustering accuracy, we consider the estimated numbers of row clusters $\hat{K}_r$, column clusters $\hat{K}_c$, and biclusters $\hat{K}_b$. In addition, we use the Rand

Table 1
Example 1: Mean, median, and standard error (SE) of $\hat{K}_r$, $\hat{K}_c$, and $\hat{K}_b$ as defined in Section 2, as well as the percentage of identifying the corresponding true numbers based on 100 replicates.

N	Method	$\hat{K}_r$				$\hat{K}_c$				$\hat{K}_b$			
		Mean	Median	SE	Per	Mean	Median	SE	Per	Mean	Median	SE	Per
30	Proposed	2.83	3.00	0.53	0.90	2.83	3.00	0.53	0.90	8.29	9.00	2.18	0.90
	bKmeans	2.76	3.00	0.64	0.66	1.13	1.00	0.46	0.05	3.09	3.00	1.34	0.03
	funHDDC	2.63	2.00	0.86	0.28	2.76	3.00	0.43	0.76	7.27	6.00	2.70	0.21
	funLBM	4.66	5.00	0.64	0.09	4.43	5.00	0.83	0.22	20.88	25.00	5.31	0.09
60	Proposed	2.91	3.00	0.43	0.93	2.90	3.00	0.41	0.94	8.61	9.00	1.74	0.93
	bKmeans	2.86	3.00	0.57	0.66	1.18	1.00	0.54	0.07	3.43	3.00	1.97	0.05
	funHDDC	2.20	2.00	0.64	0.04	2.99	3.00	0.10	0.99	6.58	6.00	1.92	0.04
	funLBM	3.42	3.00	0.64	0.66	3.24	3.00	0.55	0.82	11.15	9.00	3.31	0.55
90	Proposed	2.93	3.00	0.36	0.96	2.93	3.00	0.36	0.96	8.71	9.00	1.45	0.96
	bKmeans	2.83	3.00	0.51	0.74	1.23	1.00	0.58	0.08	3.51	3.00	1.87	0.08
	funHDDC	2.14	2.00	0.38	0.12	2.96	3.00	0.20	0.96	6.33	6.00	1.17	0.11
	funLBM	3.25	3.00	0.46	0.76	3.30	3.00	0.54	0.74	10.68	9.00	2.03	0.52

index and adjusted Rand index to assess the accuracy of clustering, including RI_r and ARI_r for row clustering, RI_c and ARI_c for column clustering, and RI_b and ARI_b for biclustering. The Rand index is defined by $\mathrm{RI} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN})$, where for example TP is the true positive count, defined as the number of sample pairs from the same cluster and assigned to the same cluster, and the other counts can be defined accordingly. As the Rand index tends to be large even under random clusterings, we also examine the adjusted Rand index defined as $\mathrm{ARI} = (\mathrm{RI} - \mathrm{E(RI)})/(\max(\mathrm{RI}) - \mathrm{E(RI)})$, which can partly correct this problem. To evaluate estimation accuracy, we examine the integrated squared error (ISE) defined as

$$\mathrm{ISE} = \frac{1}{n}\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c}}\sum_{m=1}^{n_{ij}}\left\{g_{(k_r,k_c)}(t_{i,j,m}) - \hat{g}_{i,j}(t_{i,j,m})\right\}^2.$$

We consider a total of $K_b = 9$ biclusters, which are formed by $K_r = 3$ sample (row) clusters and $K_c = 3$ covariate (column) clusters. $Y_{i,j,m} = g_{(k_r,k_c)}(t_{i,j,m}) + \epsilon_{i,j,m}$ with $t_{i,j,m}$'s, $m \in \{1, \ldots, 10\}$, equally spaced on $[0, 1]$. The nine true functional forms are $g_{(1,1)}(t) = cos(2\pi t)$, $g_{(2,1)}(t) = 1-2exp(-6t)$, $g_{(3,1)}(t) = -1.5t$, $g_{(1,2)}(t) = 1+sin(2\pi t)$, $g_{(2,2)}(t) = 2t^2$, $g_{(3,2)}(t) = t + 1$, $g_{(1,3)}(t) = 2\big(sin(2\pi t) + cos(2\pi t)\big)$, $g_{(2,3)}(t) = 1 + t^3$, and $g_{(3,3)}(t) = 2\sqrt{t} + 1$. They are also graphically presented in Fig. 1. To better mimic real data, we allow a certain proportion (ζ) of the curves from each bicluster to have 20% missing measurements. When implementing the proposed approach, we choose smoothing splines with the number of internal knots $J = 3$. We also fix $\theta = 1$ and $\tau = 3$. In what follows, under Examples 1 and 2, $N > q$, whereas under Example 3, $N = q$. Under Examples 1–3, the random errors are independent, whereas under Example 4, they are correlated. Note that under Examples 1–4, simulation results are calculated based on automatic cluster selection. Example 5 is designed to investigate the performance of these methods when the numbers of clusters are correctly prespecified. A total of 100 replicates are simulated under each setting.

Example 1. $N = 30$, 60, and 90. $q = 9$. The clusters are balanced, with each row cluster containing $N/3$ samples and each column cluster containing $q/3$ covariates. $\zeta = 0.3$. The random errors are iid $\mathcal{N}(0, 0.6^2)$.

Example 2. The settings are the same as in Example 1, except that the clusters are unbalanced. The row clusters have sizes 1:2:3, and the column clusters have sizes 2:3:4.

Example 3. Set $(N, q) = (30, 30), (39, 39), (45, 45)$, $\zeta = 0.3$ and 0.4. The rest are the same as in Example 1.

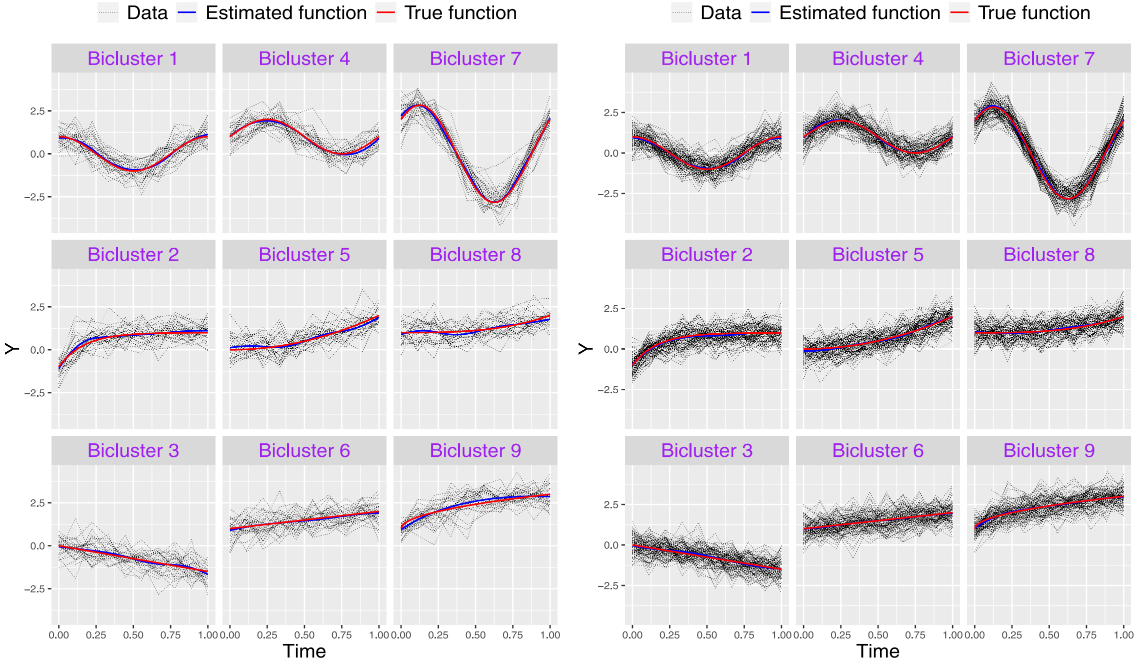
Example 4. The settings are similar to those under Example 1. The random errors are correlated with an AR(1) correlation structure, where AR stands for auto-correlation. Consider AR coefficient $\phi = 0.2$ and 0.8, representing weak and strong correlations.

Example 5. The settings are the same as those in Example 1. The difference is that the numbers of clusters are correctly prespecified instead of being selected by the BIC criterion.

Results for Example 1 are presented in Figs. 1 and 2 as well as Tables 1 and 2. More specifically, in Fig. 1, we show the true functions for all clusters as well as sample observed data and estimated functions. In Table 1, we summarize the numbers of identified row and column clusters as well as biclusters. In Table 2, we summarize the Rand and adjusted Rand index values. In Fig. 2, we present the boxplots of ISE (note that different panels have different ranges for the Y-axis). Results for Examples 2–5 are presented in the Supplementary section. Although different examples have different numerical results, overall, the advantage of the proposed approach is clearly observed. Consider for example Table 1 with

Table 2
Example 1: Mean and standard error (shown in parentheses) of RI_r, ARI_r, RI_c, ARI_c, RI_b, and ARI_b based on 100 replicates.

N	Method	RI_r	ARI_r	RI_c	ARI_c	RI_b	ARI_b
30	Proposed	0.940 (0.189)	0.911 (0.278)	0.936 (0.203)	0.910 (0.279)	0.927 (0.238)	0.909 (0.281)
	bKmeans	0.860 (0.173)	0.740 (0.290)	0.296 (0.163)	0.052 (0.194)	0.673 (0.174)	0.307 (0.167)
	funHDDC	0.744 (0.031)	0.493 (0.074)	0.940 (0.107)	0.880 (0.215)	0.889 (0.051)	0.598 (0.120)
	funLBM	0.913 (0.053)	0.786 (0.109)	0.913 (0.064)	0.746 (0.153)	0.951 (0.029)	0.708 (0.113)
60	Proposed	0.966 (0.138)	0.947 (0.208)	0.963 (0.152)	0.945 (0.212)	0.959 (0.177)	0.943 (0.216)
	bKmeans	0.887 (0.132)	0.780 (0.248)	0.316 (0.195)	0.077 (0.239)	0.704 (0.142)	0.339 (0.191)
	funHDDC	0.767 (0.021)	0.546 (0.049)	0.998 (0.025)	0.995 (0.050)	0.922 (0.014)	0.692 (0.044)
	funLBM	0.918 (0.110)	0.828 (0.221)	0.929 (0.119)	0.840 (0.257)	0.953 (0.052)	0.796 (0.198)
90	Proposed	0.978 (0.117)	0.966 (0.176)	0.975 (0.131)	0.965 (0.178)	0.971 (0.154)	0.964 (0.180)
	bKmeans	0.886 (0.134)	0.778 (0.251)	0.342 (0.226)	0.109 (0.279)	0.709 (0.152)	0.358 (0.227)
	funHDDC	0.769 (0.017)	0.551 (0.040)	0.990 (0.049)	0.980 (0.098)	0.919 (0.025)	0.686 (0.061)
	funLBM	0.909 (0.121)	0.813 (0.241)	0.908 (0.130)	0.793 (0.276)	0.944 (0.056)	0.764 (0.210)



(a) $N = 30$ **(b)** $N = 90$

Fig. 1. Example 1: Curves of observed data (black dotted), estimated (blue solid) by the proposed method, and true (red solid) functions with (a) $N = 30$ and (b) $N = 90$ for one replicate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$N = 30$. The proposed approach has the mean number of row clusters 2.83, compared to 2.76, 2.63, and 4.66 of the three alternatives. When $N = 90$, the proposed approach has the mean number of biclusters 8.71, compared to 3.51, 6.33, and 10.68 of the three alternatives. The improved clustering accuracy is further proved by the Rand index values in Table 2. For example with $N = 90$, the adjusted Rand index value for biclustering with the proposed approach is 0.964, compared to 0.358, 0.686, and 0.764 with the three alternatives. Fig. 2 shows that as N increases, estimation accuracy of the proposed approach (and two alternatives) increases. Under all three N values, the proposed approach has significantly smaller ISE values. Moreover, comparing the results of Example 5 with Example 1, we observe similar performance and that the proposed approach still performs better when the numbers of clusters are correctly prespecified.

4. Applications

Here we analyze two time-course gene expression data. Although in a sense the data characteristics are similar, the two data analyses may serve different purposes. In particular, the first dataset is “older”, which has been analyzed multiple

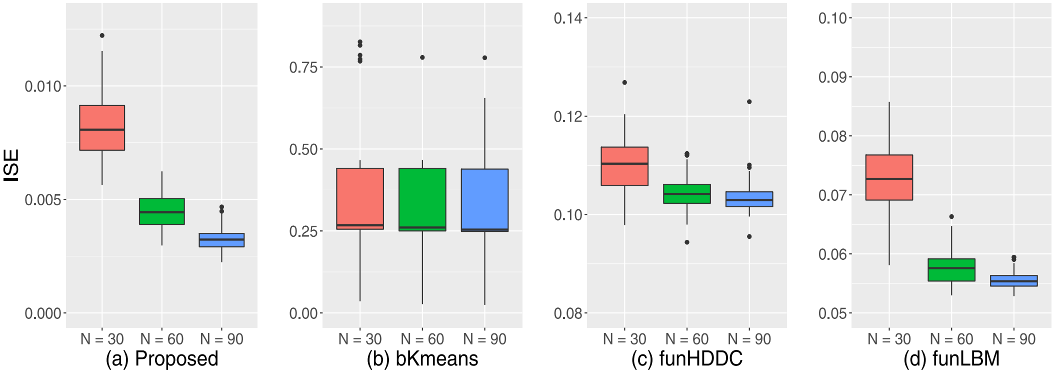


Fig. 2. Example 1: Boxplots of ISE with (a) the proposed method, (b) bKmeans, (c) funHDDC, and (d) funLBM.

times in the literature, and has a clearer sample clustering structure. In contrast, the second dataset is more recent, and its analysis may lead to a higher practical impact.

4.1. T-cell data

This data has been generated in a study of T-cell activation [31]. It is publicly available in the R package longitudinal (http://www.strimmerlab.org/software/longitudinal/) and contains two subsets: tcell.10 and tcell.34. The first subset contains measurements for 10 samples and 58 genes at 10 unequally spaced time points, $t \in \{0, 2, 4, 6, 8, 18, 24, 32, 48, 72\}$, whereas the second subset contains measurements for 34 samples and the same genes at the same time points. In [31], the distinctions between the two subsets have been noted, and they have been combined for analysis. Prior to analysis, we conduct data processing, including gene expression normalization using the method developed in [29] and linearly transforming the observed times to [0, 1], and set the knots at 0.06, 0.2, and 0.4 as well as the order as 3.

The proposed approach identifies two sample clusters, with sizes 10 and 34, which exactly match the original subset structure. The distinctions of the samples in the two subsets have been noted in [31]. As such, they are supposed to belong to different clusters. In this sense, our "finding", although as expected, is re-assuring. In addition, eight gene clusters are identified, among which there are four trivial clusters with sizes one. The four non-trivial clusters have sizes 27, 18, 5, and 4. Detailed information on the gene clusters is available from the authors. The eight non-trivial biclusters are presented in Fig. 3. Biclusters 1–4 correspond to tcell.10, and the rest correspond to tcell.34. It is observed that the estimated functions clearly differ across biclusters. The observed temporal trends are highly similar to those reported in [28], which provides support to the validity of our approach.

The three alternatives are also applied. The bKmeans approach identifies three sample clusters (with sizes 10, 17, and 17) and four gene clusters (with sizes 9, 15, 19, and 15). Compared to the proposed approach, the adjusted Rand index values are 0.441 (sample), 0.619 (gene), and 0.430 (bicluster). The funHDDC approach identifies two sample clusters (with sizes 10 and 34) and two gene clusters (with sizes 9 and 49). Compared to the proposed approach, the adjusted Rand index values are 1.000 (sample), 0.286 (gene), and 0.452 (bicluster). The funLBM approach identifies two sample clusters (with sizes 10 and 34) and six gene clusters (with sizes 9, 4, 12, 5, 18, and 10). Compared to the proposed approach, the adjusted Rand index values are 1.000 (sample), 0.586 (gene), and 0.646 (bicluster). Unlike for the simulated data, it is difficult to objectively evaluate the accuracy of clustering. However, for the proposed approach, the matching with the original sample distinction and published findings can provide a strong support, which is not shared by the alternatives.

4.2. Vaccine data

This data is generated in a relative recent study [43] and available at GEO with identifier GSE124533. The study settings have been described in detail in [43]. Briefly, it concerns with the time course of whole blood gene expressions, and the samples are healthy adults residing in an inpatient unit. The samples have been randomized into three protocols (305 A, 305B and 305C). Within each protocol, samples have been randomized to receive immunization via either vaccine or saline placebo. The treatments have been referred to as YFV and VZV (under 305 A), HBV1 and HBV3 (under 305B), and TIV and ATIV (under 305C). In this experiment, gene expression levels are measured at $t \in \{1, 2, 3, 4, 5, 7, 14, 21, 28\}$ days after immunization. A total of 43 genes have been studied, which are selected from two gene modules defined in the published literature [6,22]. Prior to analysis, gene expression normalization, rescaling of the time points (to the unit interval), and the selection of knots order are conducted in a similar way as in the previous data analysis.

Two sets of analysis are conducted. In the first set, we focus on the samples under 305 A, which contain 20 samples treated with VZV and 20 with YFV. In the second set, we pool all 122 samples from the three protocols. We note that

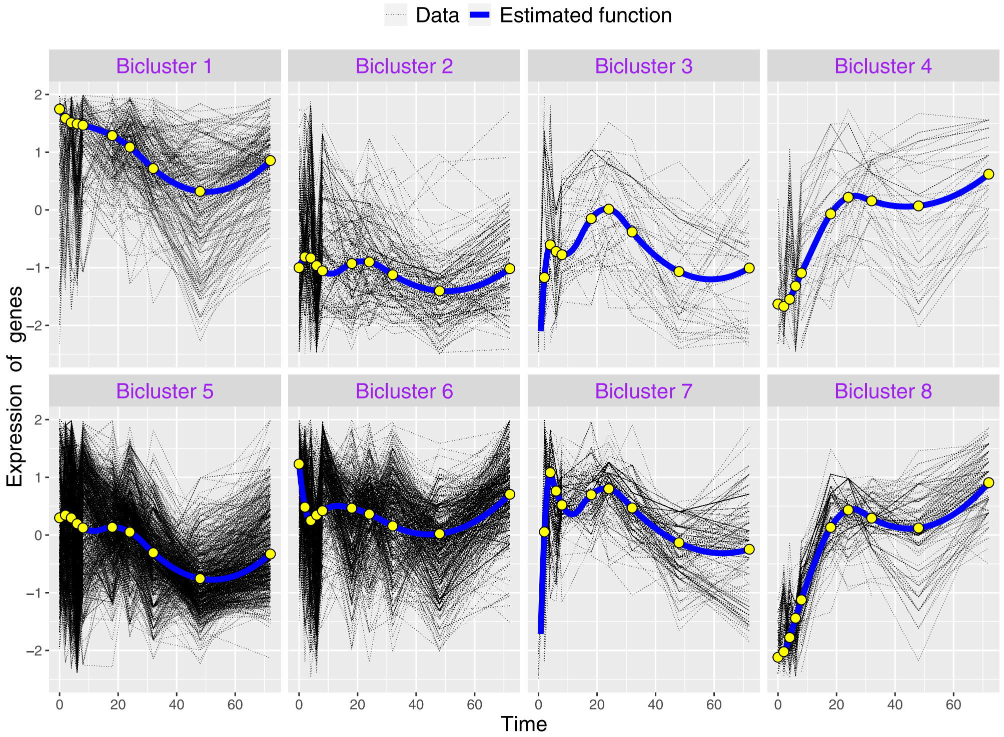


Fig. 3. Analysis of T-cell data: Curves of observed data (black dotted) and estimated functions (blue solid) for the eight non-trivial bicluster, as well as yellow points indicating the estimated values at $t \in \{0, 2, 4, 6, 8, 18, 24, 32, 48, 72\}$ by the proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

although the gene time courses have been analyzed in [43], there is insufficient attention to clustering. Complementary to the existing literature, our clustering analysis can potentially review sample heterogeneity as well as coordination among genes.

Results for the first set of analysis are presented in Fig. 5, where we observe two sample clusters and two gene clusters, leading to four biclusters. Here the two sample clusters match the VZV and YFV experimental conditions, which provides support to the validity of our analysis. The two gene clusters contain 27 and 16 members, respectively, which are very close to the module structure. Fig. 5 shows that the temporal trends of the four clusters differ significantly, with the level of variation and position of "peak" varying significantly. The observed trends are similar to those reported in [43]. We also refer to [43] for phamarcodynamic interpretations of the findings.

In the second set of analysis, we identify four sample clusters, with sizes 96, 5, 20, and 1, respectively. In what follows, we focus on the non-trivial clusters. Clusters 1 and 2 contain samples treated with VZV, HBV1, HBV3, ATIV, and TIV, and cluster 3 contains samples treated with YFV. In the original publication, there has been little attention to sample similarity/difference across protocols. Our analysis may suggest the significant difference between YFV and other treatments as well as the relative similarity of the five treatments (YFV excluded). Our analysis leads to two gene clusters, with sizes 25 and 18, respectively. This structure is again very similar to the module structure. The overall six non-trivial biclusters are shown in Fig. 4, where we observe significant across-cluster differences. Among the six patterns, biclusters 5 and 6 are similar to those observed in the first set of analysis, where biclusters 1–4 are relatively different.

The three alternatives are also applied. The bKmeans approach identifies three sample clusters (with sizes 20, 27, and 75) and two gene clusters (with sizes 26 and 17). Compared to the proposed approach, the adjusted Rand index values are 0.551 (sample), 0.907 (gene), and 0.666 (bicluster). The funHDDC approach identifies two sample clusters (with sizes 20 and 102) and three gene clusters (with sizes 26, 12 and 5). Compared to the proposed approach, the adjusted Rand index values are 0.819 (sample), 0.774 (gene), and 0.758 (bicluster). The funLBM approach identifies four sample clusters (with sizes 20, 39, 24 and 39) and two gene clusters (with sizes 20, 23). Compared to the proposed approach, the adjusted Rand index values are 0.276 (sample), 0.818 (gene), and 0.386 (bicluster).

5. Discussion

In this article, we have conducted the biclustering analysis when functions (to be exact, their realizations at discrete time points), as opposed to scalars, are present. The data structure fits time-course gene expression and other experiments. The analysis objective is considerably more complex than the biclustering analysis of scalars and one-way clustering of

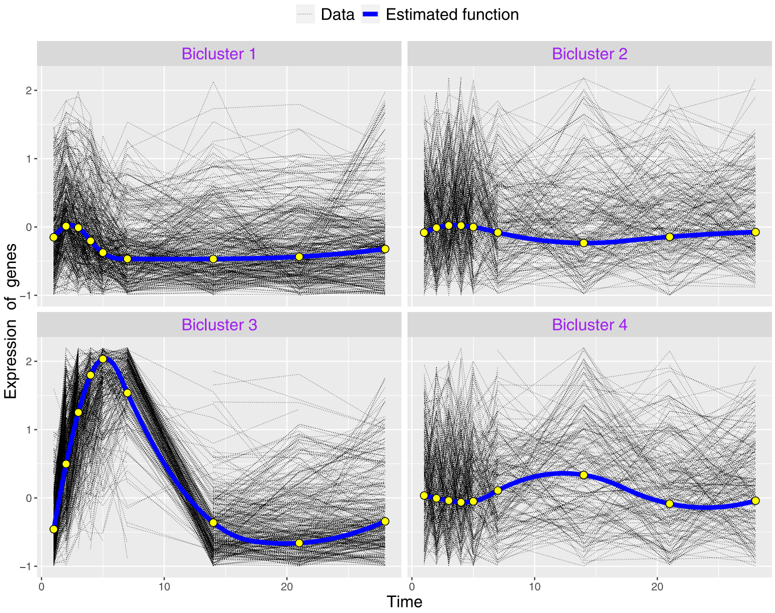


Fig. 4. Analysis of vaccine data with samples under all three protocols: Curves of observed data (black dotted) and estimated functions (blue solid) for non-trivial clusters, as well as yellow points indicating the estimated values at $t \in \{1, 2, 3, 4, 5, 7, 14, 21, 28\}$ by the proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

functions. We have developed a novel approach based on the penalized fusion technique. Methodologically, it differs significantly from the existing biclustering and fusion approaches. Theoretically, it has the much desired consistency property, making it advantageous over some of the existing alternatives that do not have theoretical support. Numerically, it has generated more accurate clustering and estimation in simulation and led to different findings in data analysis.

In our estimation, we have adopted the penalized smoothing technique. An alternative, which may be computationally simpler, is to take fewer basis functions, with which we can eliminate the smoothness penalty. Theoretically and numerically, we expect similar performance. The fusion technique involves pairwise differences/penalties, which may incur higher computational cost when N and/or q are large. In our simulation, we have considered moderate values, which match our data analysis. It will be of interest to develop computationally more scalable approaches/algorithms, for example via model averaging. This is beyond our scope and will be postponed to the future. In data analysis, findings with certain support have been made. In the literature, most existing studies are on the "static" functionalities of genes. It will be important to further understand the dynamics of gene expressions and more solidly interpret the findings.

CRediT authorship contribution statement

Kuangnan Fang: Methodology, Formal analysis Writing – original draft. **Yuanxing Chen:** Data curation, Investigation, Software, Writing – original draft. **Shuangge Ma:** Conceptualization, Methodology, Writing – review & editing. **Qingzhao Zhang:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Acknowledgments

We thank the Editor-in-Chief, Managing Editor, and two reviewers for insightful comments and suggestions. This work was supported by the National Natural Science Foundation of China (11971404, 72071169), Humanity and Social Science Youth Foundation of Ministry of Education of China (19YJC910010), Basic Scientific Project 71988101 of National Science Foundation of China, 111 Project (B13028), National Institutes of Health (CA204120), and National Science Foundation (1916251).

Appendix A. Proofs

Proof of Proposition 1. By the definitions of $\mathbf{H}_r^{(m+1)}$ and $\mathbf{H}_c^{(m+1)}$, for any $\mathbf{H}_r$ and $\mathbf{H}_c$, we have

$$L_\theta\left(\boldsymbol{\beta}^{(m+1)}, \mathbf{H}_r^{(m+1)}, \mathbf{H}_c^{(m+1)}, \boldsymbol{\Lambda}_r^{(m)}, \boldsymbol{\Lambda}_c^{(m)}\right) \le L_\theta\left(\boldsymbol{\beta}^{(m+1)}, \mathbf{H}_r, \mathbf{H}_c, \boldsymbol{\Lambda}_r^{(m)}, \boldsymbol{\Lambda}_c^{(m)}\right).$$

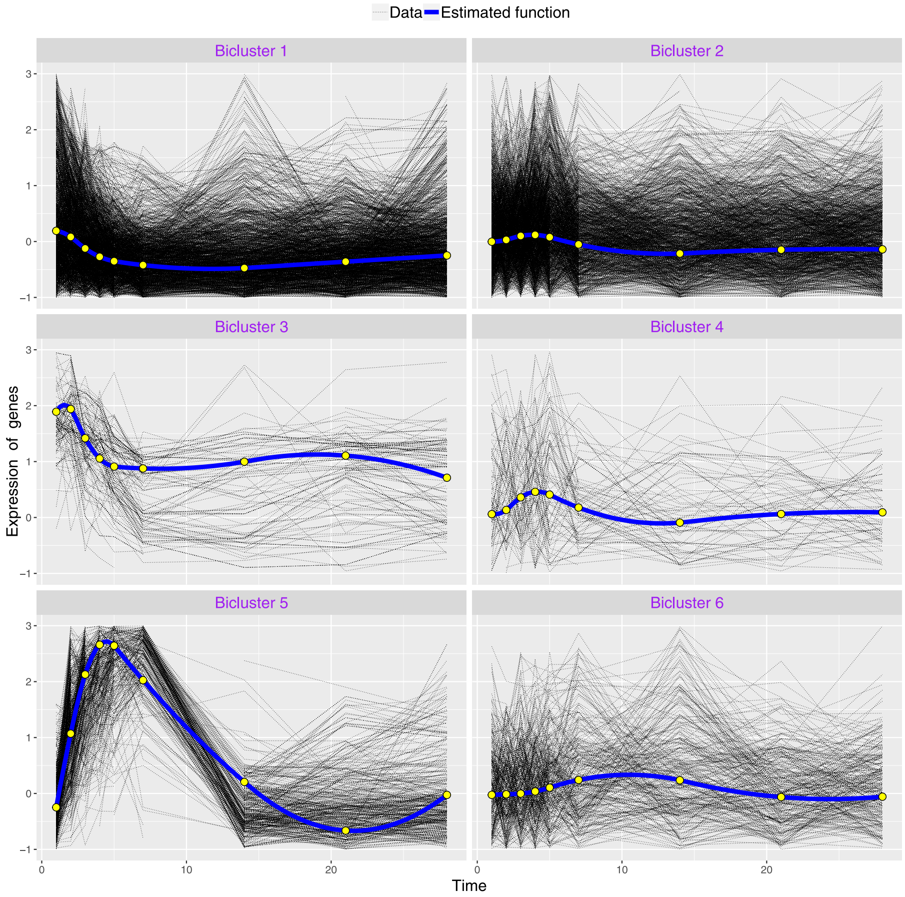


Fig. 5. Analysis of vaccine data with samples under 305A: Curves of observed data (black dotted) and estimated functions (blue solid), as well as yellow points indicating the estimated values at $t \in \{1,2,3,4,5,7,14,21,28\}$ by the proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Let $\Xi(\boldsymbol{\beta}^{(m+1)}) = \left\{(\mathbf{H}_r,\mathbf{H}_c) \,:\, \mathbf{B}_r\boldsymbol{\beta}^{(m+1)} - \mathrm{vec}(\mathbf{H}_r) = \mathbf{0}, \mathbf{B}_c\boldsymbol{\beta}^{(m+1)} - \mathrm{vec}(\mathbf{H}_c) = \mathbf{0}\right\}$ and $\boldsymbol{P} = \sum_{\delta\in\Delta^{(r)}} p_\tau(\|\boldsymbol{\eta}_\delta^{(r)}\|_2,\gamma_2) + \sum_{\delta\in\Delta^{(c)}} p_\tau(\|\boldsymbol{\eta}_\delta^{(c)}\|_2,(N/q)^{1/2}\gamma_2)$. We can define

$$f^{(m+1)} = \inf_{\Xi(\boldsymbol{\beta}^{(m+1)})}\left\{\frac{1}{2}\|\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+1)}\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^{(m+1)\top}\mathbf{M}\boldsymbol{\beta}^{(m+1)} + \boldsymbol{P}\right\} = \inf_{\Xi(\boldsymbol{\beta}^{(m+1)})} L_\theta\left(\boldsymbol{\beta}^{(m+1)},\mathbf{H}_r,\mathbf{H}_c,\boldsymbol{\Lambda}_r^{(m)},\boldsymbol{\Lambda}_c^{(m)}\right),$$

and then $L_\theta\left(\boldsymbol{\beta}^{(m+1)},\mathbf{H}_r^{(m+1)},\mathbf{H}_c^{(m+1)},\boldsymbol{\Lambda}_r^{(m)},\boldsymbol{\Lambda}_c^{(m)}\right) \le f^{(m+1)}$.

For any integer n, we have $\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m+n-1)}) = \mathrm{vec}(\boldsymbol{\Lambda}_r^{(m)}) + \theta\sum_{i=1}^{n-1}\left[\mathrm{vec}(\mathbf{H}_r^{(m+i)}) - \mathbf{B}_r\boldsymbol{\beta}^{(m+i)}\right]$ and $\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m+n-1)}) = \mathrm{vec}(\boldsymbol{\Lambda}_c^{(m)}) + \theta\sum_{i=1}^{n-1}\left[\mathrm{vec}(\mathbf{H}_c^{(m+i)}) - \mathbf{B}_c\boldsymbol{\beta}^{(m+i)}\right]$, and then

$$\begin{aligned}
&L_\theta\left(\boldsymbol{\beta}^{(m+n)},\mathbf{H}_r^{(m+n)},\mathbf{H}_c^{(m+n)},\boldsymbol{\Lambda}_r^{(m+n-1)},\boldsymbol{\Lambda}_c^{(m+n-1)}\right)\\
&= \frac{1}{2}\|\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+n)}\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^{(m+n)\top}\mathbf{M}\boldsymbol{\beta}^{(m+n)} + \boldsymbol{P} + \left\{\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m)}) + \theta\sum_{i=1}^{n-1}\left[\mathrm{vec}(\mathbf{H}_r^{(m+i)}) - \mathbf{B}_r\boldsymbol{\beta}^{(m+i)}\right]\right\}^\top\\
&\quad\times\left[\mathrm{vec}(\mathbf{H}_r^{(m+n)}) - \mathbf{B}_r\boldsymbol{\beta}^{(m+n)}\right]\\
&\quad+\left\{\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m)}) + \theta\sum_{i=1}^{n-1}\left[\mathrm{vec}(\mathbf{H}_c^{(m+i)}) - \mathbf{B}_c\boldsymbol{\beta}^{(m+i)}\right]\right\}^\top\left[\mathrm{vec}(\mathbf{H}_c^{(m+n)}) - \mathbf{B}_c\boldsymbol{\beta}^{(m+n)}\right] + \frac{\theta}{2}\left\|\mathrm{vec}(\mathbf{H}_r^{(m+n)} - \mathbf{B}_r\boldsymbol{\beta}^{(m+n)})\right\|_2^2\\
&\quad+\frac{\theta}{2}\left\|\mathrm{vec}(\mathbf{H}_c^{(m+n)} - \mathbf{B}_c\boldsymbol{\beta}^{(m+n)})\right\|_2^2 \le f^{(m+n)}.
\end{aligned}$$

Since the augmented Lagrangian function $L_\theta\Big(\boldsymbol{\beta}, \mathbf{H}_r, \mathbf{H}_c, \boldsymbol{\Lambda}_r \boldsymbol{\Lambda}_c\Big)$ is differentiable with respect to $\boldsymbol{\beta}$ and is convex with respect to each $\boldsymbol{\eta}_\delta^{(r)}$ and $\boldsymbol{\eta}_\delta^{(c)}$. By Theorem 4.1 of [38], there exists a limit point of $(\boldsymbol{\beta}^{(m)}, \mathbf{H}_r^{(m)}, \mathbf{H}_c^{(m)})$, denoted by $(\boldsymbol{\beta}^*, \mathbf{H}_r^*, \mathbf{H}_c^*)$. Then we have

$$f^* = \lim_{m\to\infty} f^{(m+1)} = \lim_{m\to\infty} f^{(m+n)} = \inf_{\Xi(\boldsymbol{\beta}^*)}\left\{\frac{1}{2}\|\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^*\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^{*\top}\mathbf{M}\boldsymbol{\beta}^* + \boldsymbol{P}\right\}.$$

For all $t \geq 0$, we have

$$\begin{aligned}
&\lim_{m\to\infty} L_\theta\Big(\boldsymbol{\beta}^{(m+n)}, \mathbf{H}_r^{(m+n)}, \mathbf{H}_c^{(m+n)}, \boldsymbol{\Lambda}_r^{(m+n-1)}, \boldsymbol{\Lambda}_c^{(m+n-1)}\Big)\\
&= \frac{1}{2}\|\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^*\|_2^2 + \frac{1}{2}\gamma_1\boldsymbol{\beta}^{*\top}\mathbf{M}\boldsymbol{\beta}^* \ + \boldsymbol{P} + \lim_{m\to\infty}\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m)})^\top\Big[\mathrm{vec}(\mathbf{H}_r^*) - \mathbf{B}_r\boldsymbol{\beta}^*\Big] + (n-\frac{1}{2})\theta\Big\|\mathrm{vec}(\mathbf{H}_r^*) - \mathbf{B}_r\boldsymbol{\beta}^*\Big\|_2^2\\
&\quad + \lim_{m\to\infty}\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m)})^\top\Big[\mathrm{vec}(\mathbf{H}_c^*) - \mathbf{B}_c\boldsymbol{\beta}^*\Big] + (n-\frac{1}{2})\theta\Big\|\mathrm{vec}(\mathbf{H}_c^*) - \mathbf{B}_c\boldsymbol{\beta}^*\Big\|_2^2 \leq f^*.
\end{aligned}$$

Thus

$$\lim_{m\to\infty}\Big\|\mathbf{r}_r^{(m+1)}\Big\|_2^2 = \Big\|\mathbf{B}_r\boldsymbol{\beta}^* - \mathrm{vec}(\mathbf{H}_r^*)\Big\|_2^2 = 0, \quad \lim_{m\to\infty}\Big\|\mathbf{r}_c^{(m+1)}\Big\|_2^2 = \Big\|\mathbf{B}_c\boldsymbol{\beta}^* - \mathrm{vec}(\mathbf{H}_c^*)\Big\|_2^2 = 0.$$

Besides, by the definition of $\boldsymbol{\beta}^{(m+1)}$, we have that

$$\begin{aligned}
&\partial L_\theta\Big(\boldsymbol{\beta}^{(m+1)}, \mathbf{H}_r^{(m+1)}, \mathbf{H}_c^{(m+1)}, \boldsymbol{\Lambda}_r^{(m)}, \boldsymbol{\Lambda}_c^{(m)}\Big)/\partial\boldsymbol{\beta}\\
&= -\mathbf{U}^\top(\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+1)}) + \gamma_1\mathbf{M}\boldsymbol{\beta}^{(m+1)} - \theta\mathbf{B}_r^\top\Big[\mathrm{vec}(\mathbf{H}_r^{(m)}) + \mathrm{vec}(\mathbf{\Lambda}_r^{(m)})/\theta - \mathbf{B}_r\boldsymbol{\beta}^{(m+1)}\Big]\\
&\quad - \theta\mathbf{B}_c^\top\Big[\mathrm{vec}(\mathbf{H}_c^{(m)}) + \mathrm{vec}(\mathbf{\Lambda}_c^{(m)})/\theta - \mathbf{B}_c\boldsymbol{\beta}^{(m+1)}\Big]\\
&= -\mathbf{U}^\top(\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+1)}) + \gamma_1\mathbf{M}\boldsymbol{\beta}^{(m+1)} - \mathbf{B}_r^\top\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m)}) - \theta\mathbf{B}_r^\top\Big[\mathrm{vec}(\mathbf{H}_r^{(m)}) - \mathbf{B}_r\boldsymbol{\beta}^{(m+1)}\Big]\\
&\quad - \mathbf{B}_c^\top\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m)}) - \theta\mathbf{B}_c^\top\Big[\mathrm{vec}(\mathbf{H}_c^{(m)}) - \mathbf{B}_c\boldsymbol{\beta}^{(m+1)}\Big]\\
&= -\mathbf{U}^\top(\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+1)}) + \gamma_1\mathbf{M}\boldsymbol{\beta}^{(m+1)} - \mathbf{B}_r^\top\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m+1)}) + \theta\mathbf{B}_r^\top\Big[\mathrm{vec}(\mathbf{H}_r^{(m+1)}) - \mathrm{vec}(\mathbf{H}_r^{(m)})\Big]\\
&\quad - \mathbf{B}_c^\top\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m+1)}) + \theta\mathbf{B}_c^\top\Big[\mathrm{vec}(\mathbf{H}_c^{(m+1)}) - \mathrm{vec}(\mathbf{H}_c^{(m)})\Big] = 0.
\end{aligned}$$

Then we can obtain

$$\mathbf{s}_r^{(m+1)} + \mathbf{s}_c^{(m+1)} = \mathbf{U}^\top(\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+1)}) - \gamma_1\mathbf{M}\boldsymbol{\beta}^{(m+1)} + \mathbf{B}_r^\top\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m+1)}) + \mathbf{B}_c^\top\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m+1)}).$$

By $\Big\|\mathbf{B}_r\boldsymbol{\beta}^* - \mathrm{vec}(\mathbf{H}_r^*)\Big\|_2^2 = 0$ and $\Big\|\mathbf{B}_c\boldsymbol{\beta}^* - \mathrm{vec}(\mathbf{H}_c^*)\Big\|_2^2 = 0$, we have

$$\begin{aligned}
&\lim_{m\to\infty}\partial L_\theta\Big(\boldsymbol{\beta}^{(m+1)}, \mathbf{H}_r^{(m+1)}, \mathbf{H}_c^{(m+1)}, \boldsymbol{\Lambda}_r^{(m)}, \boldsymbol{\Lambda}_c^{(m)}\Big)/\partial\boldsymbol{\beta}\\
&= -\mathbf{U}^\top(\mathbf{Y}-\mathbf{U}\boldsymbol{\beta}^{(m+1)}) + \gamma_1\mathbf{M}\boldsymbol{\beta}^{(m+1)} - \mathbf{B}_r^\top\mathrm{vec}(\boldsymbol{\Lambda}_r^{(m+1)}) - \mathbf{B}_c^\top\mathrm{vec}(\boldsymbol{\Lambda}_c^{(m+1)}) = \mathbf{0}.
\end{aligned}$$

Therefore $\lim_{m\to\infty}\mathbf{s}_r^{(m+1)} + \mathbf{s}_c^{(m+1)} = \mathbf{0}$. □

Let $|\mathcal{G}_{k_r,k_c}^{(r,c)*}| = \sum_{(i,j)\in\mathcal{G}_{k_r,k_c}^{(r,c)}} n_{i,j}$ and $n_m = \max_{i\in\{1,\ldots,N\}, j\in\{1,\ldots,q\}} n_{i,j} < \infty$. Then $|\mathcal{G}_{k_r,k_c}^{(r,c)}| \leq |\mathcal{G}_{k_r,k_c}^{(r,c)*}| \leq n_m|\mathcal{G}_{k_r,k_c}^{(r,c)}|$. Denote the number of internal knots as J and then $J = p - d$. Recall that $b = \min_{(k_r,k_c)\neq(k_r',k_c')}\|g_{(k_r,k_c)}^* - g_{(k_r',k_c')}^*\|$.

Lemma 1. *Under Condition (C1), there exists a spline approximation $\boldsymbol{\alpha}_{k_r,k_c}^{*\top}U_p(t)$ of the true function $g_{(k_r,k_c)}^*(t)$ for $k_r \in \{1,\ldots,K_r\}$ and $k_c \in \{1,\ldots,K_c\}$, such that*

$$\sup_{t\in\mathcal{T}}|g_{(k_r,k_c)}^*(t) - \boldsymbol{\alpha}_{k_r,k_c}^{*\top}U_p(t)| = O(J^{-\kappa}).$$

Proof. Lemma 1 follows from Corollary 6.21 of [34]. This lemma has been used in a number of studies that involve spline expansion [25,42]. We omit the proof here. □

Lemma 2. *Under Conditions (C1)–(C3) and $b \gg J^{-\kappa}$, there exists a constant $C_2 > 0$ such that for all $(k_r, k_c) \neq (k_r', k_c')$, such that*

$$\|\alpha_{k_r,k_c}^* - \alpha_{k_r',k_c'}^*\|_2 \geq \frac{1}{2}C_2^{-1/2}b,$$

when N and q are sufficiently large.

Proof. By the triangular inequality, we have

$$\|(\alpha^*_{k_r,k_c} - \alpha^*_{k_r',k_c'})^\top U_p\| \geq \|g^*_{(k_r,k_c)} - g^*_{(k_r',k_c')}\| - \|g^*_{(k_r,k_c)} - \alpha^{*\top}_{k_r,k_c} U_p\| - \|g^*_{(k_r',k_c')} - \alpha^{*\top}_{k_r',k_c'} U_p\|. \tag{A.1}$$

Besides, by Theorem 5.4.2 of [13], Condition (C2), and the definition of the rescaled B-spline basis, for any vector $\boldsymbol{\alpha}'_{p\times 1}$, there exists a constant $C_2 > 0$ such that

$$\|\boldsymbol{\alpha}'^\top U_p\|^2 \leq C_2 \|\boldsymbol{\alpha}'\|_2^2. \tag{A.2}$$

Combining (A.1), (A.2), and Lemma 1, we have

$$\begin{aligned}\|\alpha^*_{k_r,k_c} - \alpha^*_{k_r',k_c'}\|_2 &\geq C_2^{-1/2}\left\{\|g^*_{(k_r,k_c)} - g^*_{(k_r',k_c')}\| - \|g^*_{(k_r,k_c)} - \alpha^{*\top}_{k_r,k_c} U_p\| - \|g^*_{(k_r,k_c)} - \alpha^{*\top}_{k_r,k_c} U_p\|\right\} \\ &\geq C_2^{-1/2}(b - 2M_2 J^{-\kappa}) > C_2^{-1/2}(b - 2 \times \frac{1}{4} b) = \frac{1}{2} C_2^{-1/2} b,\end{aligned}$$

where the third inequality is obtained when N and q are sufficiently large since $b \gg J^{-\kappa}$. □

Lemma 3 (*Bernstein's Inequality, Lemma 2.2.11 in [39]*). *For independent random variables $Y_1, \dots, Y_n$ with means 0 and $E|Y_i|^m \leq m! M^{m-2} v_i/2$ for some constants M, v_i, and every $m \geq 2$,*

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2\exp\left\{-\frac{1}{2}\frac{x^2}{v + Mx}\right\},$$

where $v = v_1 + \cdots + v_n$.

Proof of Theorem 1. Given $\hat{\boldsymbol{\beta}}^{or} \in \mathcal{M}_G$, when the true block memberships $\mathcal{G}^{(r,c)}_{1,1}, \dots, G^{(r,c)}_{K_r,K_c}$ are known, the oracle estimators for all $\boldsymbol{\beta}_{i,j}$'s are the same if $(i.j) \in \mathcal{G}^{(r,c)}_{k_r,k_c}$. Thus we can explore the properties of $\hat{\boldsymbol{\beta}}^{or}$ by examining the properties of the oracle common coefficient vector $\hat{\boldsymbol{\alpha}}^{or} = (\hat{\boldsymbol{\alpha}}^{or\top}_{1,1}, \dots, \hat{\boldsymbol{\alpha}}^{or\top}_{k_r,k_c}, \dots, \hat{\boldsymbol{\alpha}}^{or\top}_{K_r,K_c})^\top$, which is defined as

$$\hat{\boldsymbol{\alpha}}^{or} = \arg\min_{\boldsymbol{\alpha}} \sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c} \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c}),$$

and

$$\hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c}) = \frac{1}{2}\|\mathbf{Y}_{(k_r,k_c)} - \mathbf{U}_{(k_r,k_c)}\boldsymbol{\alpha}_{k_r,k_c}\|_2^2 + \gamma_1 |\mathcal{G}^{(r,c)}_{k_r,k_c}| \boldsymbol{\alpha}^\top_{k_r,k_c}\mathbf{D}\boldsymbol{\alpha}_{k_r,k_c},$$

where $\mathbf{Y}_{(k_r,k_c)} = \mathrm{vec}\{Y_{i,j}, (i,j) \in \mathcal{G}_{k_r,k_c}\}$, $\mathbf{U}_{(k_r,k_c)} = (\mathbf{U}^\top_{i,j}, (i,j) \in \mathcal{G}_{k_r,k_c})^\top$. The corresponding true B-spline coefficient vector is denoted by $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}^{*\top}_{1,1}, \dots, \boldsymbol{\alpha}^{*\top}_{k_r,k_c}, \dots, \boldsymbol{\alpha}^{*\top}_{K_r,K_c})^\top$. Note that

$$\left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c}} - \left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \boldsymbol{\alpha}^*_{k_r,k_c}} = \left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}\partial \boldsymbol{\alpha}^\top_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \bar{\boldsymbol{\alpha}}_{k_r,k_c}} (\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c} - \boldsymbol{\alpha}^*_{k_r,k_c}),$$

where $\bar{\boldsymbol{\alpha}}_{k_r,k_c}$ is between $\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c}$ and $\boldsymbol{\alpha}^*_{k_r,k_c}$. Then we have

$$\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c} - \boldsymbol{\alpha}^*_{k_r,k_c} = -\left(\left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}\partial \boldsymbol{\alpha}^\top_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \bar{\boldsymbol{\alpha}}_{k_r,k_c}}\right)^{-1} \left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \boldsymbol{\alpha}^*_{k_r,k_c}}.$$

Hence

$$\|\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c} - \boldsymbol{\alpha}^*_{k_r,k_c}\|_2 \leq \left\| |\mathcal{G}^{(r,c)*}_{k_r,k_c}| \left(\left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}\partial \boldsymbol{\alpha}^\top_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \bar{\boldsymbol{\alpha}}_{k_r,k_c}}\right)^{-1}\right\|_2 \left\| |\mathcal{G}^{(r,c)*}_{k_r,k_c}|^{-1} \left.\frac{\partial \hat{L}^{or}(\boldsymbol{\alpha}_{k_r,k_c})}{\partial \boldsymbol{\alpha}_{k_r,k_c}}\right|_{\boldsymbol{\alpha}_{k_r,k_c} = \boldsymbol{\alpha}^*_{k_r,k_c}}\right\|_2 := A^{(1)}_{k_r,k_c} \times A^{(2)}_{k_r,k_c}. \tag{A.3}$$

By Lemma A.8 of [41], Conditions (C1) and (C2), we can derive that there exists a constant $C_3 > 0$ such that for any $1 \leq k_r \leq K_r$, $1 \leq k_c \leq K_c$,

$$P(A^{(1)}_{k_r,k_c} \leq C_3) = P\left(\left\|\frac{\mathbf{U}^\top_{(k_r,k_c)}\mathbf{U}_{(k_r,k_c)}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|} + \frac{\gamma_1 |\mathcal{G}^{(r,c)}_{k_r,k_c}|\mathbf{D}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\right\|_2 \leq C_3\right) \geq 1 - p/(Nq). \tag{A.4}$$

Besides, note that

$$\begin{aligned} A^{(2)}_{k_r,k_c} &= \left\| -\frac{\mathbf{U}^\top_{(k_r,k_c)}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\left(\mathbf{Y}_{(k_r,k_c)} - \mathbf{g}^*_{(k_r,k_c)} + \mathbf{g}^*_{(k_r,k_c)} - \mathbf{U}_{(k_r,k_c)}\boldsymbol{\alpha}^*_{k_r,k_c}\right) + \gamma_1 \frac{|\mathcal{G}^{(r,c)}_{k_r,k_c}|}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\mathbf{D}\boldsymbol{\alpha}^*_{k_r,k_c}\right\|_2 \\ &\le \left\| \frac{\mathbf{U}^\top_{(k_r,k_c)}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\boldsymbol{\epsilon}_{k_r,k_c}\right\|_2 + \left\| \frac{\mathbf{U}^\top_{(k_r,k_c)}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}(\mathbf{g}^*_{(k_r,k_c)} - \mathbf{U}_{(k_r,k_c)}\boldsymbol{\alpha}^*_{k_r,k_c})\right\|_2 + \left\| \gamma_1 \frac{|\mathcal{G}^{(r,c)}_{k_r,k_c}|}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\mathbf{D}\boldsymbol{\alpha}^*_{k_r,k_c}\right\|_2 := B^{(1)}_{k_r,k_c} + B^{(2)}_{k_r,k_c} + B^{(3)}_{k_r,k_c}. \end{aligned} \tag{A.5}$$

Since the rescaled B-spline values are finite, there exists constant $M_1 > 0$ such that $U_{l,p}(t) \le M_1$ for $l \in \{1, \dots, p\}$. Let $\mathbf{U}_{(i,j)\cdot l}$ denote the lth column of $\mathbf{U}_{(i,j)}$, and we verify the condition of Lemma 3 by Condition (C5)

$$\mathrm{E}|\mathbf{U}^\top_{(i,j)\cdot l}\boldsymbol{\epsilon}_{i,j}|^m \le \ \mathrm{E}\Big(|\mathbf{U}^\top_{(i,j)\cdot l}\mathbf{U}_{(i,j)\cdot l}|^{m/2} \cdot |\boldsymbol{\epsilon}^\top_{i,j}\boldsymbol{\epsilon}_{i,j}|^{m/2}\Big) \le (F^{-1}M_1)^m m!\mathrm{E}\Big(\exp\big\{F|n^{-1}_{i,j}\boldsymbol{\epsilon}^\top_{i,j}\boldsymbol{\epsilon}_{i,j}|^{1/2}\big\}\Big) \le (F^{-1}M_1)^m m! c_2.$$

Applying Lemma 3, we have

$$P\Bigg(\Bigg|\sum_{(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c}} \mathbf{U}^\top_{(i,j)\cdot l}\boldsymbol{\epsilon}_{i,j}\Bigg| > x\Bigg) \le 2\exp\left\{-\frac{1}{2}\frac{x^2}{v + F^{-1}M_1 x}\right\}, \tag{A.6}$$

where $v = \sum_{(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c}} v_{i,j}$ and $v_{i,j} = 2F^{-2}M_1^2 c_2$.

Let $\mathbf{U}_{(k_r,k_c)\cdot l}$ denote the lth column of $\mathbf{U}_{(k_r,k_c)}$. For some constant $0 < C < \infty$, combining Condition (C5) and (A.6), we have

$$\begin{aligned} &P\ \Big(\ \big\||\mathcal{G}^{(r,c)*}_{k_r,k_c}|^{-1}\mathbf{U}^\top_{(k_r,k_c)}\boldsymbol{\epsilon}_{k_r,k_c}\big\|_\infty > CF^{-1}M_1\big(\log(Nq)/|\mathcal{G}^{(r,c)*}_{k_r,k_c}|\big)^{1/2}\ \Big) \\ &\le \sum_{l=1}^{p} P\Big(|\mathbf{U}^\top_{(k_r,k_c)\cdot l}\boldsymbol{\epsilon}_{k_r,k_c}| > CF^{-1}M_1\big(\log(Nq)|\mathcal{G}^{(r,c)*}_{k_r,k_c}|\big)^{1/2}\Big) \\ &= \sum_{l=1}^{p} P\Bigg(\Bigg|\sum_{(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c}} \mathbf{U}^\top_{(i,j)\cdot l}\boldsymbol{\epsilon}_{i,j}\Bigg| > CF^{-1}M_1\big(\log(Nq)|\mathcal{G}^{(r,c)*}_{k_r,k_c}|\big)^{1/2}\Bigg) \\ &\le 2p\exp\left\{-\frac{1}{2}\frac{C^2F^{-2}M_1^2\big(\log(Nq)|\mathcal{G}^{(r,c)*}_{k_r,k_c}|\big)}{2F^{-2}M_1^2c_2|\mathcal{G}^{(r,c)}_{k_r,k_c}| + CF^{-2}M_1^2\big(\log(Nq)|\mathcal{G}^{(r,c)*}_{k_r,k_c}|\big)^{1/2}}\right\} \le 2p\exp\Big\{-\log(Nq)\Big\} \le \ 2p/Nq. \end{aligned}$$

Hence, we have that with probability at least $1 - 2p/(Nq)$,

$$B^{(1)}_{k_r,k_c} \le CF^{-1}M_1\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{k_r,k_c}|\big)^{1/2}. \tag{A.7}$$

By Lemma 1, there exists a constant $M_2 > 0$ such that

$$B^{(2)}_{k_r,k_c} \le p^{1/2}\left\|\frac{\mathbf{U}^\top_{(k_r,k_c)}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}(\mathbf{g}^*_{(k_r,k_c)} - \mathbf{U}_{(k_r,k_c)}\boldsymbol{\alpha}^*_{k_r,k_c})\right\|_\infty \le p^{1/2}\left\|\frac{\mathbf{U}^\top_{(k_r,k_c)}}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\right\|_\infty \left\|(\mathbf{g}^*_{(k_r,k_c)} - \mathbf{U}_{(k_r,k_c)}\boldsymbol{\alpha}^*_{k_r,k_c})\right\|_\infty \le M_1M_2p^{1/2}J^{-\kappa}. \tag{A.8}$$

In addition,

$$B^{(3)}_{k_r,k_c} \le \gamma_1\frac{|\mathcal{G}^{(r,c)}_{k_r,k_c}|}{|\mathcal{G}^{(r,c)*}_{k_r,k_c}|}\|\boldsymbol{\alpha}^*_{k_r,k_c}\|_2\|\mathbf{D}\|_2 \le p^{1/2}\gamma_1\|\boldsymbol{\alpha}^*_{k_r,k_c}\|_\infty\|\mathbf{D}\|_2. \tag{A.9}$$

Thus by (A.5), (A.7), (A.8), and (A.9), for any $1 \le k_r \le K_r$, $1 \le k_c \le K_c$, with probability at least $1 - 2p/(Nq)$,

$$A^{(2)}_{k_r,k_c} \le CF^{-1}M_1\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2} + M_1M_2p^{1/2}J^{-\kappa} + \max_{k_r,k_c}\|\boldsymbol{\alpha}^*_{k_r,k_c}\|_\infty\|\mathbf{D}\|_2\gamma_1p^{1/2}.$$

By Condition (C1) and $\gamma_1 = o(|\mathcal{G}^{(r,c)}_{min}|^{-1/2})$, when N and q are sufficiently large, we have

$$p^{1/2}J^{-\kappa} \ll \big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}, \quad p^{1/2}\gamma_1 \ll \big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}.$$

Hence, for any $1 \le k_r \le K_r$, $1 \le k_c \le K_c$, with probability at least $1 - 2p/(Nq)$,

$$A^{(2)}_{k_r,k_c} \le C_4\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2},$$

where C_4 is a large constant. Together with (A.3) and (A.4), for any $1 \le k_r \le K_r$, $1 \le k_c \le K_c$,

$$\begin{aligned} P\Big(\|\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c} - \boldsymbol{\alpha}^*_{k_r,k_c}\|_2 \le C_3C_4\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}\Big) &\ge 1 - P(A^{(1)}_{k_r,k_c} > C_3) - P\Big(A^{(2)}_{k_r,k_c} > C_4\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}\Big) \\ &\ge 1 - 3p/(Nq). \end{aligned}$$

By the Bonferroni's inequality, we have

$$
\begin{aligned}
&P\Big(\sup_{1\le k_r\le K_r,1\le k_c\le K_c}\|\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c}-\boldsymbol{\alpha}^{*}_{k_r,k_c}\|_2\le C_3C_4\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}\Big)\\
&\ge 1-\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}P\Big(\|\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c}-\boldsymbol{\alpha}^{*}_{k_r,k_c}\|_2> C_3C_4\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}\Big)\ge 1-3K_rK_cp/(Nq).
\end{aligned}
$$

By Lemma 1 and (A.2), we have

$$
\begin{aligned}
\|\hat{g}^{or}_{(k_r,k_c)}-g^{*}_{(k_r,k_c)}\| &= \|\hat{\boldsymbol{\alpha}}^{or\top}_{k_r,k_c}U_p-\boldsymbol{\alpha}^{*\top}_{k_r,k_c}U_p+\boldsymbol{\alpha}^{*\top}_{k_r,k_c}U_p-\mathbf{g}^{*}_{(k_r,k_c)}\|\le\|(\hat{\boldsymbol{\alpha}}^{or}_{k_r,k_c}-\boldsymbol{\alpha}^{*}_{k_r,k_c})^\top U_p\|+\|\boldsymbol{\alpha}^{*\top}_{k_r,k_c}U_p-\mathbf{g}^{*}_{(k_r,k_c)}\|\\
&\le C_2^{1/2}C_3C_4\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}+M_2J^{-\kappa}\le (C_2^{1/2}C_3C_4+M_2/2)\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}\\
&= C^{*}\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2},
\end{aligned}
$$

where $C^*=\max\{C_3C_4, C_2^{1/2}C_3C_4+M_2/2\}$. That is,

$$
P\Big(\sup_{1\le k_r\le K_r,1\le k_c\le K_c}\|\hat{g}^{or}_{(k_r,k_c)}-g^{*}_{(k_r,k_c)}\|\le\psi\Big)\ge 1-3K_rK_cp/(Nq),
$$

where $\psi=C^*\big(p\log(Nq)/|\mathcal{G}^{(r,c)}_{min}|\big)^{1/2}$. □

Proof of Theorem 2. Let $\rho_1(t)=\gamma_2^{-1}p_\tau(t,\gamma_2)$ and $\rho_2(t)=((N/q)^{1/2}\gamma_2)^{-1}p_\tau(t,(N/q)^{1/2}\gamma_2)$. Define

$$
\begin{aligned}
Q(\boldsymbol{\beta}) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q}\Big(\|\mathbf{Y}_{i,j}-\mathbf{U}_{i,j}\boldsymbol{\beta}_{i,j}\|_2^2+\gamma_1\boldsymbol{\beta}_{i,j}^\top\mathbf{D}\boldsymbol{\beta}_{i,j}\Big),\\
Pen(\boldsymbol{\beta}) &= \gamma_2\sum_{(i_1,i_2)\in\Delta^{(r)}}\rho_1(\|\boldsymbol{\beta}^{(r)}_{i_1}-\boldsymbol{\beta}^{(r)}_{i_2}\|_2)+(N/q)^{1/2}\gamma_2\sum_{(j_1,j_2)\in\Delta^{(c)}}\rho_2(\|\boldsymbol{\beta}^{(c)}_{j_1}-\boldsymbol{\beta}^{(c)}_{j_2}\|_2),\\
Q^{\mathcal{G}}(\boldsymbol{\alpha}) &= \frac{1}{2}\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\Big(\|\mathbf{Y}_{(k_r,k_c)}-\mathbf{U}_{(k_r,k_c)}\boldsymbol{\alpha}_{k_r,k_c}\|_2^2+\gamma_1|\mathcal{G}^{(r,c)}_{k_r,k_c}|\boldsymbol{\alpha}^\top_{k_r,k_c}\mathbf{D}\boldsymbol{\alpha}_{k_r,k_c}\Big),\\
Pen^{\mathcal{G}}(\boldsymbol{\alpha}) &= \gamma_2\sum_{k_r<k_r'}|\mathcal{G}^{(r)}_{k_r}||\mathcal{G}^{(r)}_{k_r'}|\rho_1(\|\boldsymbol{\alpha}^{(r)}_{k_r}-\boldsymbol{\alpha}^{(r)}_{k_r'}\|_2)+(N/q)^{1/2}\gamma_2\sum_{k_c<k_c'}|\mathcal{G}^{(c)}_{k_c}||\mathcal{G}^{(c)}_{k_c'}|\rho_2(\|\boldsymbol{\alpha}^{(c)}_{k_c}-\boldsymbol{\alpha}^{(c)}_{k_c'}\|_2),
\end{aligned}
$$

where $\boldsymbol{\alpha}^{(r)}_{k_r}=(\boldsymbol{\alpha}^{(r)\top}_{k_r,1},\ldots,\boldsymbol{\alpha}^{(r)\top}_{k_r,q})^\top$ with $\boldsymbol{\alpha}^{(r)}_{k_r,j}=\boldsymbol{\alpha}_{k_r,k}$ if $j\in\mathcal{G}^{(c)}_k$, $\boldsymbol{\alpha}^{(c)}_{k_c}=(\boldsymbol{\alpha}^{(c)\top}_{1,k_c},\ldots,\boldsymbol{\alpha}^{(c)\top}_{N,k_c})^\top$ with $\boldsymbol{\alpha}^{(c)}_{i,k_c}=\boldsymbol{\alpha}_{k,k_c}$ if $i\in\mathcal{G}^{(r)}_k$. Let

$$
L(\boldsymbol{\beta})=Q(\boldsymbol{\beta})+Pen(\boldsymbol{\beta}),\quad L^{\mathcal{G}}(\boldsymbol{\alpha})=Q^{\mathcal{G}}(\boldsymbol{\alpha})+Pen^{\mathcal{G}}(\boldsymbol{\alpha}).
$$

We define two mappings, $\widetilde{T}:\mathcal{M}_{\mathcal{G}}\to\widetilde{\mathcal{M}}_{\mathcal{G}}$ and $\widehat{T}:\mathbb{R}^{Nqp}\to\widehat{\mathcal{M}}_{\mathcal{G}}$, and the two subspaces are defined by

$$
\begin{aligned}
\widetilde{\mathcal{M}}_{\mathcal{G}} &= \Big\{\boldsymbol{\alpha}\in\mathbb{R}^{K_rK_cp}:\boldsymbol{\alpha}_{k_r,k_c}=\boldsymbol{\beta}_{i,j},\text{ for any }(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c},\ 1\le k_r\le K_r,\ 1\le k_c\le K_c\Big\},\\
\widehat{\mathcal{M}}_{\mathcal{G}} &= \Big\{\boldsymbol{\alpha}\in\mathbb{R}^{K_rK_cp}:\boldsymbol{\alpha}_{k_r,k_c}=|\mathcal{G}^{(r,c)}_{k_r,k_c}|^{-1}\sum_{(i,j)\in\mathcal{G}^{(r,c)}_{k_r,k_c}}\boldsymbol{\beta}_{i,j},\ 1\le k_r\le K_r,\ 1\le k_c\le K_c\Big\}.
\end{aligned}
$$

For every $\boldsymbol{\beta}\in\mathcal{M}_{\mathcal{G}}$, we have $Pen(\boldsymbol{\beta})=Pen^{\mathcal{G}}(\widetilde{T}(\boldsymbol{\beta}))$, and for every $\boldsymbol{\alpha}\in\widetilde{\mathcal{M}}_{\mathcal{G}}$, we have $Pen(\widetilde{T}^{-1}(\boldsymbol{\alpha}))=Pen^{\mathcal{G}}(\boldsymbol{\alpha})$. Hence

$$
L(\boldsymbol{\beta})=L^{\mathcal{G}}(\widetilde{T}(\boldsymbol{\beta})),\quad L^{\mathcal{G}}(\boldsymbol{\alpha})=L(\widetilde{T}^{-1}(\boldsymbol{\alpha})). \tag{A.10}
$$

Consider the neighborhood of $\boldsymbol{\beta}^*$:

$$
\Theta=\Big\{\boldsymbol{\beta}\in\mathbb{R}^{Nqp}:\sup_{1\le i\le N,1\le j\le q}\|\boldsymbol{\beta}_{i,j}-\boldsymbol{\beta}^{*}_{i,j}\|_2\le\psi\Big\}.
$$

By the result in Theorem 1, there is an event E_1 such that on E_1,

$$
\sup_{1\le i\le N,1\le j\le q}\|\hat{\boldsymbol{\beta}}^{or}_{i,j}-\boldsymbol{\beta}^{*}_{i,j}\|_2\le\psi,
$$

and $P(E_1^C)\le 3K_rK_cp/(Nq)$. Hence $\hat{\boldsymbol{\beta}}^{or}\in\Theta$ on E_1. For any $\boldsymbol{\beta}\in\mathbb{R}^{Nqp}$, let $\tilde{\boldsymbol{\beta}}=\widetilde{T}^{-1}(\widehat{T}(\boldsymbol{\beta}))$. Inspired by [27], we show that $\hat{\boldsymbol{\beta}}^{or}$ is a strictly local minimizer of objective function (3) with probability tending to 1 through the following two steps:

(i) On E_1, $L(\tilde{\boldsymbol{\beta}})>L(\hat{\boldsymbol{\beta}}^{or})$ for any $\boldsymbol{\beta}\in\Theta$ and $\tilde{\boldsymbol{\beta}}\ne\hat{\boldsymbol{\beta}}^{or}$.

(ii) There is an event E_2 such that $P(E_2^C) \le c_2/(Nq)$. On $E_1 \cap E_2$, there is a neighborhood of $\hat{\boldsymbol{\beta}}^{or}$, denoted by Θ', such that $L(\boldsymbol{\beta}) \ge L(\tilde{\boldsymbol{\beta}})$ for any $\boldsymbol{\beta} \in \Theta' \cap \Theta$ for sufficiently large N and q.

Therefore, by the results in (i) and (ii), we have $L(\boldsymbol{\beta}) > L(\hat{\boldsymbol{\beta}}^{or})$ for any $\boldsymbol{\beta} \in \Theta' \cap \Theta$ and $\tilde{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^{or}$, so that $\hat{\boldsymbol{\beta}}^{or}$ is a strictly local minimizer of $L(\boldsymbol{\beta})$ on $E_1 \cap E_2$ with $P(E_1 \cap E_2) \ge 1 - 3K_rK_pp/(Nq) - c_2/(Nq)$ for sufficiently large N and q.

Firstly, we prove the result in (i) . Let $\widehat{T}(\boldsymbol{\beta}) = \boldsymbol{\alpha} = (\boldsymbol{\alpha}_{1,1}^\top, \ldots, \boldsymbol{\alpha}_{K_r,K_c}^\top)^\top$ and $\boldsymbol{\alpha}_{k_r}^{(r)*} = (\boldsymbol{\beta}_{i,1}^{(r)*\top}, \ldots, \boldsymbol{\beta}_{i,q}^{(r)*\top})^\top$ for $i \in \mathcal{G}_{k_r}^{(r)}$. Since

$$\|\boldsymbol{\alpha}_{k_r}^{(r)} - \boldsymbol{\alpha}_{k_r'}^{(r)}\|_2 \ge \|\boldsymbol{\alpha}_{k_r}^{(r)*} - \boldsymbol{\alpha}_{k_r'}^{(r)*}\|_2 - 2 \sup_{1\le k_r \le K_r} \|\boldsymbol{\alpha}_{k_r}^{(r)} - \boldsymbol{\alpha}_{k_r}^{(r)*}\|_2,$$

and

$$\begin{aligned}\sup_{1\le k_r\le K_r} \| \boldsymbol{\alpha}_{k_r}^{(r)} - \boldsymbol{\alpha}_{k_r}^{(r)*} \|_2^2 &= \sup_{1\le k_r\le K_r} \left\{ \sum_{k_c=1}^{K_c} |\mathcal{G}_{k_c}^{(c)}| \cdot \left\| \sum_{i\in\mathcal{G}_{k_r}^{(r)}} \sum_{j\in\mathcal{G}_{k_c}^{(c)}} \boldsymbol{\beta}_{i,j}/(|\mathcal{G}_{k_r}^{(r)} \| \mathcal{G}_{k_c}^{(c)}|) - \boldsymbol{\alpha}_{k_r,k_c}^* \right\|_2^2 \right\} \\ &\le \sup_{1\le k_r\le K_r} |\mathcal{G}_{k_r}^{(r)}|^{-1} \sum_{k_c=1}^{K_c} \sum_{i\in\mathcal{G}_{k_r}^{(r)}} \sum_{j\in\mathcal{G}_{k_c}^{(c)}} \| \boldsymbol{\beta}_{i,j} - \boldsymbol{\beta}_{i,j}^* \|_2^2 \le q \sup_{1\le i\le N, 1\le j\le q} \|\boldsymbol{\beta}_{i,j} - \boldsymbol{\beta}_{i,j}^*\|_2^2\end{aligned} \tag{A.11}$$

by Lemma 2, for any $k_r \ne k_r'$

$$\begin{aligned}\|\boldsymbol{\alpha}_{k_r}^{(r)} - \boldsymbol{\alpha}_{k_r'}^{(r)}\|_2 &\ge \frac{1}{2}|\mathcal{G}_{min}^{(c)}|^{1/2} C_2^{-1/2} b - 2q^{1/2} \sup_{1\le i\le N, 1\le j\le q} \|\boldsymbol{\beta}_{i,j} - \boldsymbol{\beta}_{i,j}^*\|_2 \\ &\ge \frac{1}{2}|\mathcal{G}_{min}^{(c)}|^{1/2} C_2^{-1/2} b - 2q^{1/2} C_3 \psi > a\gamma_2.\end{aligned}$$

The last inequality follows from the assumption that $|\mathcal{G}_{min}^{(c)}|^{1/2} b \gg \gamma_2 \gg \left(pq\right)^{1/2} \log(Nq)/\min\{|\mathcal{G}_{min}^{(r)}|, |\mathcal{G}_{min}^{(c)}|\} \gg q^{1/2}\psi$. Similarly, for any $k_c \ne k_c'$, we have

$$\begin{aligned}\|\boldsymbol{\alpha}_{k_c}^{(c)} - \boldsymbol{\alpha}_{k_c'}^{(c)}\|_2 &\ge \frac{1}{2}|\mathcal{G}_{min}^{(r)}|^{1/2} C_2^{-1/2} b - 2N^{1/2} \sup_{1\le i\le N, 1\le j\le q} \|\boldsymbol{\beta}_{i,j} - \boldsymbol{\beta}_{i,j}^*\|_2 \\ &\ge \frac{1}{2}|\mathcal{G}_{min}^{(r)}|^{1/2} C_2^{-1/2} b - 2N^{1/2} C_3 \psi > a(N/q)^{1/2}\gamma_2.\end{aligned}$$

Hence by Condition (C4), $Pen^{\mathcal{G}}(\widehat{T}(\boldsymbol{\beta})) = C_{pen}$, a constant, and hence $L^{\mathcal{G}}(\widehat{T}(\boldsymbol{\beta})) = Q^{\mathcal{G}}(\widehat{T}(\boldsymbol{\beta})) + C_{pen}$ for all $\boldsymbol{\beta} \in \Theta$. Since $\hat{\boldsymbol{\alpha}}^{or}$ is the unique global minimizer of $Q^{\mathcal{G}}(\boldsymbol{\alpha})$, $Q^{\mathcal{G}}(\widehat{T}(\boldsymbol{\beta})) > Q^{\mathcal{G}}(\hat{\boldsymbol{\alpha}}^{or})$ for all $\widehat{T}(\boldsymbol{\beta}) \ne \hat{\boldsymbol{\alpha}}^{or}$, and thus $L^{\mathcal{G}}(\widehat{T}(\boldsymbol{\beta})) > L^{\mathcal{G}}(\hat{\boldsymbol{\alpha}}^{or})$ for all $\widehat{T}(\boldsymbol{\beta}) \ne \hat{\boldsymbol{\alpha}}^{or}$. By (A.10), we have $L^{\mathcal{G}}(\widehat{T}(\boldsymbol{\beta})) = L(\tilde{\boldsymbol{\beta}})$ and $L^{\mathcal{G}}(\hat{\boldsymbol{\alpha}}^{or}) = L(\hat{\boldsymbol{\beta}}^{or})$. Therefore $L(\tilde{\boldsymbol{\beta}}) > L(\hat{\boldsymbol{\beta}}^{or})$ for all $\tilde{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^{or}$, and the result (i) is proved.

Next we prove result (ii). For a positive sequence ν_n, let

$$\Theta' = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{Nqp} : \sup_{1\le i\le N} \|\boldsymbol{\beta}_i^{(r)} - \hat{\boldsymbol{\beta}}_i^{(r)or}\|_2 \le \nu_n, \quad \sup_{1\le j\le q} \|\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}}_j^{(c)or}\|_2 \le \nu_n \right\},$$

$$Pen^r(\boldsymbol{\beta}) = \gamma_2 \sum_{(i_1,i_2)\in\Delta^{(r)}} \rho_1(\|\boldsymbol{\beta}_{i_1}^{(r)} - \boldsymbol{\beta}_{i_2}^{(r)}\|_2), \quad Pen^c(\boldsymbol{\beta}) = (N/q)^{1/2}\gamma_2 \sum_{(j_1,j_2)\in\Delta^{(c)}} \rho_2(\|\boldsymbol{\beta}_{j_1}^{(c)} - \boldsymbol{\beta}_{j_2}^{(c)}\|_2),$$

and $Pen(\boldsymbol{\beta}) = Pen^r(\boldsymbol{\beta}) + Pen^c(\boldsymbol{\beta})$. For $\boldsymbol{\beta} \in \Theta' \cap \Theta$, by Taylor's expansion, we have

$$L(\boldsymbol{\beta}) - L(\tilde{\boldsymbol{\beta}}) = \Omega_1 + \Omega_2 + \Omega_3, \tag{A.12}$$

where

$$\Omega_1 = \sum_{i=1}^{N}\sum_{j=1}^{q} \Big[-\mathbf{U}_{i,j}^\top(\mathbf{Y}_{i,j} - \mathbf{U}_{i,j}\bar{\boldsymbol{\beta}}_{i,j}) + \gamma_1 \mathbf{D}\bar{\boldsymbol{\beta}}_{i,j} \Big]^\top (\boldsymbol{\beta}_{i,j} - \tilde{\boldsymbol{\beta}}_{i,j}),$$

$$\Omega_2 = \sum_{i=1}^{N} \left(\frac{\partial Pen^r(\bar{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_i^{(r)}} \bigg|_{\boldsymbol{\beta}_i^{(r)} = \bar{\boldsymbol{\beta}}_i^{(r)}} \right)^\top (\boldsymbol{\beta}_i^{(r)} - \tilde{\boldsymbol{\beta}}_i^{(r)}), \quad \Omega_3 = \sum_{j=1}^{q} \left(\frac{\partial Pen^c(\bar{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_j^{(c)}} \bigg|_{\boldsymbol{\beta}_j^{(c)} = \bar{\boldsymbol{\beta}}_j^{(c)}} \right)^\top (\boldsymbol{\beta}_j^{(c)} - \tilde{\boldsymbol{\beta}}_j^{(c)}),$$

with $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\beta}}_{1,1}^\top, \ldots, \bar{\boldsymbol{\beta}}_{N,q}^\top)^\top$ and $\bar{\boldsymbol{\beta}}_{i,j} = s\boldsymbol{\beta}_{i,j} + (1-s)\tilde{\boldsymbol{\beta}}_{i,j}$ for some $s \in (0, 1)$.

Firstly, we have

$$\begin{aligned}\Omega_2 &= \gamma_2 \sum_{i_1<i_2} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2^{-1}(\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)})^\top(\boldsymbol{\beta}_{i_1}^{(r)} - \tilde{\boldsymbol{\beta}}_{i_1}^{(r)})\\ &\quad + \gamma_2 \sum_{i_1>i_2} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2^{-1}(\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)})^\top(\boldsymbol{\beta}_{i_1}^{(r)} - \tilde{\boldsymbol{\beta}}_{i_1}^{(r)})\\ &= \gamma_2 \sum_{i_1<i_2} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2^{-1}(\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)})^\top[(\boldsymbol{\beta}_{i_1}^{(r)} - \tilde{\boldsymbol{\beta}}_{i_1}^{(r)}) - (\boldsymbol{\beta}_{i_2}^{(r)} - \tilde{\boldsymbol{\beta}}_{i_2}^{(r)})].\end{aligned}$$

When $i_1, i_2 \in \mathcal{G}_{k_r}^{(r)}$, $\tilde{\boldsymbol{\beta}}_{i_1} = \tilde{\boldsymbol{\beta}}_{i_2}$. Thus

$$\begin{aligned}\Omega_2 &= \gamma_2 \sum_{k_r=1}^{K_r} \sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}, i_1<i_2} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2^{-1}(\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)})^\top(\boldsymbol{\beta}_{i_1}^{(r)} - \boldsymbol{\beta}_{i_2}^{(r)})\\ &\quad + \gamma_2 \sum_{k_r<k_r'} \sum_{i_1\in\mathcal{G}_{k_r}^{(r)}, i_2\in\mathcal{G}_{k_r'}^{(r)}} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2^{-1}(\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)})^\top[(\boldsymbol{\beta}_{i_1}^{(r)} - \tilde{\boldsymbol{\beta}}_{i_1}^{(r)}) - (\boldsymbol{\beta}_{i_2}^{(r)} - \tilde{\boldsymbol{\beta}}_{i_2}^{(r)})].\end{aligned}$$

As shown in Theorem 1, $\sup_i \|\tilde{\boldsymbol{\beta}}_i^{(r)} - \boldsymbol{\beta}_i^{(r)*}\|_2^2 = \sup_{k_r} \|\boldsymbol{\alpha}_{k_r}^{(r)} - \boldsymbol{\alpha}_{k_r}^{(r)*}\|_2^2 \le q\psi^2$. Since $\bar{\boldsymbol{\beta}}_i^{(r)} = s\boldsymbol{\beta}_i^{(r)} + (1-s)\tilde{\boldsymbol{\beta}}_i^{(r)}$, $\sup_i \|\bar{\boldsymbol{\beta}}_i^{(r)} - \boldsymbol{\beta}_i^{(r)*}\|_2 \le sq^{1/2}\psi + (1-s)q^{1/2}\psi = q^{1/2}\psi$. For $k_r \neq k_r'$, $i_1 \in \mathcal{G}_{k_r}^{(r)}$, $i_2 \in \mathcal{G}_{k_r'}^{(r)}$, we have

$$\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2 \ge \min_{i_1\in\mathcal{G}_{k_r}^{(r)}, i_2\in\mathcal{G}_{k_r'}^{(r)}} \|\boldsymbol{\beta}_{i_1}^{(r)*} - \boldsymbol{\beta}_{i_2}^{(r)*}\|_2 - 2\max_i \|\bar{\boldsymbol{\beta}}_i^{(r)} - \boldsymbol{\beta}_i^{(r)*}\|_2 \ge \frac{1}{2}|\mathcal{G}_{min}^{(c)}|^{1/2} C_2^{-1/2} b - 2q^{1/2}\psi > a\gamma_2,$$

and thus $\rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2) = 0$. Therefore,

$$\begin{aligned}\Omega_2 &= \gamma_2 \sum_{k_r=1}^{K_r} \sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}, i_1<i_2} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\boldsymbol{\beta}_{i_1}^{(r)} - \boldsymbol{\beta}_{i_2}^{(r)}\|_2\\ &\ge \gamma_2 \sum_{k_r=1}^{K_r} \sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}, i_1<i_2} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2) q^{-1/2} \sum_{j=1}^{q} \|\boldsymbol{\beta}_{i_1,j} - \boldsymbol{\beta}_{i_2,j}\|_2\\ &= \gamma_2 q^{-1/2} \sum_{k_r=1}^{K_r} \sum_{k_c=1}^{K_c} \sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}, i_1<i_2} \sum_{j\in\mathcal{G}_{k_c}^{(c)}} \rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2)\|\boldsymbol{\beta}_{i_1,j} - \boldsymbol{\beta}_{i_2,j}\|_2.\end{aligned}$$

Similarly to (A.11), $\sup_i \|\tilde{\boldsymbol{\beta}}_i^{(r)} - \hat{\boldsymbol{\beta}}_i^{(r)or}\|_2 \le \nu_n$ and $\sup_i \|\boldsymbol{\beta}_i^{(r)} - \hat{\boldsymbol{\beta}}_i^{(r)or}\|_2 \le \nu_n$. Then we have

$$\sup_{i_1<i_2} \|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2 \le 2\sup_i \|\bar{\boldsymbol{\beta}}_i^{(r)} - \tilde{\boldsymbol{\beta}}_i^{(r)}\|_2 \le 2\sup_i \|\boldsymbol{\beta}_i^{(r)} - \tilde{\boldsymbol{\beta}}_i^{(r)}\|_2 \le 2\Big(\sup_i \|\boldsymbol{\beta}_i^{(r)} - \hat{\boldsymbol{\beta}}_i^{(r)or}\|_2 + \sup_i \|\tilde{\boldsymbol{\beta}}_i^{(r)} - \hat{\boldsymbol{\beta}}_i^{(r)or}\|_2\Big) \le 4\nu_n.$$

Hence $\rho_1'(\|\bar{\boldsymbol{\beta}}_{i_1}^{(r)} - \bar{\boldsymbol{\beta}}_{i_2}^{(r)}\|_2) \ge \rho_1'(4\nu_n)$ by the concavity of $\rho(\cdot)$. As a result,

$$\Omega_2 \ge \gamma_2 q^{-1/2} \sum_{k_r=1}^{K_r} \sum_{k_c=1}^{K_c} \sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}, i_1<i_2} \sum_{j\in\mathcal{G}_{k_c}^{(c)}} \rho_1'(4\nu_n)\|\boldsymbol{\beta}_{i_1,j} - \boldsymbol{\beta}_{i_2,j}\|_2. \tag{A.13}$$

Next we consider Ω_3. Similarly to the derivation of (A.13), we can derive

$$\Omega_3 \ge \gamma_2 q^{-1/2} \sum_{k_r=1}^{K_r} \sum_{k_c=1}^{K_c} \sum_{j_1,j_2\in\mathcal{G}_{k_c}^{(c)}, j_1<j_2} \sum_{i\in\mathcal{G}_{k_r}^{(r)}} \rho_2'(4\nu_n)\|\boldsymbol{\beta}_{i,j_1} - \boldsymbol{\beta}_{i,j_2}\|_2. \tag{A.14}$$

Lastly for Ω_1, we have

$$\Omega_1 = -\sum_{i=1}^{N} \sum_{j=1}^{q} \mathbf{w}_{i,j}^\top(\boldsymbol{\beta}_{i,j} - \tilde{\boldsymbol{\beta}}_{i,j}) = -\sum_{k_r=1}^{K_r} \sum_{k_c=1}^{K_c} \sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}} \sum_{j_1,j_2\in\mathcal{G}_{k_c}^{(c)}} \frac{\mathbf{w}_{i_1,j_1}^\top(\boldsymbol{\beta}_{i_1,j_1} - \boldsymbol{\beta}_{i_2,j_2})}{|\mathcal{G}_{k_r,k_c}^{(r,c)}|},$$

and

$$\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}}\sum_{j_1,j_2\in\mathcal{G}_{k_c}^{(c)}}\frac{|\mathbf{w}_{i_1,j_1}^\top(\boldsymbol{\beta}_{i_1,j_1}-\boldsymbol{\beta}_{i_2,j_2})|}{|\mathcal{G}_{k_r,k_c}^{(r,c)}|}\leq\sup_{i,j}\|\mathbf{w}_{i,j}\|_2\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)}}\sum_{j_1,j_2\in\mathcal{G}_{k_c}^{(c)}}\frac{\|\boldsymbol{\beta}_{i_1,j_1}-\boldsymbol{\beta}_{i_2,j_2}\|_2}{|\mathcal{G}_{k_r,k_c}^{(r,c)}|}$$

$$\leq 2\sup_{i,j}\|\mathbf{w}_{i,j}\|_2\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)},i_1<i_2}\sum_{j\in\mathcal{G}_{k_c}^{(c)}}\frac{\|\boldsymbol{\beta}_{i_1,j}-\boldsymbol{\beta}_{i_2,j}\|_2}{|\mathcal{G}_{k_r}^{(r)}|}+2\sup_{i,j}\|\mathbf{w}_{i,j}\|_2\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{j_1,j_2\in\mathcal{G}_{k_c}^{(c)},j_1<j_2}\sum_{i\in\mathcal{G}_{k_r}^{(r)}}\frac{\|\boldsymbol{\beta}_{i,j_1}-\boldsymbol{\beta}_{i,j_2}\|_2}{|\mathcal{G}_{k_c}^{(c)}|},$$

where $\mathbf{w}_{i,j}=\mathbf{U}_{i,j}^\top(\mathbf{Y}_{i,j}-\mathbf{U}_{i,j}\bar{\boldsymbol{\beta}}_{i,j})-\gamma_1\mathbf{D}\bar{\boldsymbol{\beta}}_{i,j}$. Note that

$$\sup_{i,j}\|\mathbf{w}_{i,j}\|_2\leq\sup_{i,j}\|\mathbf{U}_{i,j}^\top(\mathbf{g}_{i,j}^*-\mathbf{U}_{i,j}\boldsymbol{\beta}_{i,j}^*)\|_2+\sup_{i,j}\|(\mathbf{U}_{i,j}^\top\mathbf{U}_{i,j}+\gamma_1\mathbf{D})(\boldsymbol{\beta}_{i,j}^*-\bar{\boldsymbol{\beta}}_{i,j})\|_2+\sup_{i,j}\|\gamma_1\mathbf{D}\boldsymbol{\beta}_{i,j}^*\|_2+\sup_{i,j}\|\mathbf{U}_{i,j}^\top\boldsymbol{\epsilon}_{i,j}\|_2.$$

By Lemma 1, $\sup_{i,j}\|\mathbf{U}_{i,j}^\top(\mathbf{g}_{i,j}^*-\mathbf{U}_{i,j}\boldsymbol{\beta}_{i,j}^*)\|_2\leq n_mM_1M_2p^{1/2}J^{-\kappa}$. Moreover, $\sup_{i,j}\|(\mathbf{U}_{i,j}^\top\mathbf{U}_{i,j}+\gamma_1\mathbf{D})(\boldsymbol{\beta}_{i,j}^*-\bar{\boldsymbol{\beta}}_{i,j})\|_2\leq(n_m^{1/2}p^{1/2}M_1+\gamma_1\|\mathbf{D}\|_2)\psi$, $\sup_{i,j}\|\gamma_1\mathbf{D}\boldsymbol{\beta}_{i,j}^*\|_2\leq p^{1/2}\gamma_1\|\mathbf{D}\|_2\|\boldsymbol{\beta}^*\|_\infty$. With the Bonferroni's inequality, Markov's inequality, and Condition (C5), we have

$$P\Big(\sup_{i,j}\|\mathbf{U}_{(i,j)}^\top\boldsymbol{\epsilon}_{i,j}\|_2>2n_{i,j}F^{-1}M_1p^{1/2}\log(Nq)\Big)\leq\sum_{i=1}^{N}\sum_{j=1}^{q}P\Big(\|\mathbf{U}_{(i,j)}^\top\boldsymbol{\epsilon}_{i,j}\|_2>2n_{i,j}F^{-1}M_1p^{1/2}\log(Nq)\Big)$$

$$\leq\sum_{i=1}^{N}\sum_{j=1}^{q}P\Big(Fn_{i,j}^{-1/2}\|\boldsymbol{\epsilon}_{i,j}\|_2>2\log(Nq)\Big)\leq c_2/(Nq).$$

Together with Conditions (C1) and (C3), we have

$$\sup_{i,j}\|\mathbf{w}_{i,j}\|_2=O(p^{1/2}\log(Nq))\tag{A.15}$$

holds with probability at least $1-c_2/(Nq)$. Let $\nu_n=o(1)$, then $\rho_1'(4\nu_n)\to 1$ and $\rho_2'(4\nu_n)\to 1$. Since $\gamma_2\gg(pq)^{1/2}\log(Nq)/\min\{|\mathcal{G}_{min}^{(r)}|,|\mathcal{G}_{min}^{(c)}|\}$, then by (A.12)–(A.15)

$$L(\boldsymbol{\beta})-L(\tilde{\boldsymbol{\beta}})=\Omega_1+\Omega_2+\Omega_3\geq\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{i_1,i_2\in\mathcal{G}_{k_r}^{(r)},i_1<i_2}\sum_{j\in\mathcal{G}_{k_c}^{(c)}}\left[\gamma_2q^{-1/2}\rho_1'(4\nu_n)-\frac{2\sup_{i,j}\|\mathbf{w}_{i,j}\|_2}{|\mathcal{G}_{k_r}^{(r)}|}\right]\|\boldsymbol{\beta}_{i_1,j}-\boldsymbol{\beta}_{i_2,j}\|_2$$

$$+\sum_{k_r=1}^{K_r}\sum_{k_c=1}^{K_c}\sum_{j_1,j_2\in\mathcal{G}_{k_c}^{(c)},j_1<j_2}\sum_{i\in\mathcal{G}_{k_r}^{(r)}}\left[\gamma_2q^{-1/2}\rho_2'(4\nu_n)-\frac{2\sup_{i,j}\|\mathbf{w}_{i,j}\|_2}{|\mathcal{G}_{k_c}^{(c)}|}\right]\|\boldsymbol{\beta}_{i,j_1}-\boldsymbol{\beta}_{i,j_2}\|_2\geq 0$$

holds with probability at least $1-c_2/(Nq)$, which completes the proof of result (ii). □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104874. The Supplementary section contains additional tables and figures for Examples 2–5.

References

[1] C. Abraham, P.A. Cornillon, E. Matzner-Løber, N. Molinari, Unsupervised curve clustering using B-splines, Scand. J. Stat. 30 (3) (2003) 581–595.

[2] G. Aneiros, R. Cao, R. Fraiman, C. Genest, P. Vieu, Recent advances in functional data analysis and high-dimensional statistics, J. Multivariate Anal. 170 (2019) 3–9.

[3] G. Biau, L. Devroye, G. Lugosi, On the performance of clustering in hilbert spaces, IEEE Trans. Inform. Theory 54 (2) (2008) 781–790.

[4] C. Bouveyron, L. Bozzi, J. Jacques, F.-X. Jollois, The functional latent block model for the co-clustering of electricity consumption curves, J. R. Stat. Soc. Ser. C. Appl. Stat. 67 (4) (2018) 897–915.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.

[6] D. Chaussabel, C. Quinn, J. Shen, P. Patel, C. Glaser, N. Baldwin, D. Stichweh, D. Blankenship, L. Li, A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus, Immunity 29 (1) (2008) 150–164.

[7] J. Chen, S. Zhang, Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data, Bioinformatics 32 (11) (2016) 1724–1732.

[8] E.C. Chi, K. Lange, Splitting methods for convex clustering, J. Comput. Graph. Statist. 24 (4) (2015) 994–1013.

[9] J.-M. Chiou, P.-L. Li, Functional clustering and identifying substructures of longitudinal data, J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (4) (2007) 679–699.

[10] J.-M. Chiou, P.-L. Li, Correlation-based functional clustering via subspace projection, J. Amer. Statist. Assoc. 103 (484) (2008) 1684–1692.

[11] W. Chu, R. Li, R. Matthew, Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data, Ann. Appl. Stat. 10 (2) (2016) 596–617.

[12] N. Coffey, J. Hinde, E. Holian, Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data, Comput. Statist. Data Anal. 71 (2014) 14–29.
[13] R.A. DeVore, G.G. Lorentz, Constructive Approximation: Polynomials and Splines Approximation, Springer-Verlag, Berlin, 1993.
[14] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (456) (2001) 1348–1360.
[15] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, J. Multivariate Anal. 146 (146) (2016) 1–6.
[16] B.P. Hejblum, J. Skinner, R. Thiébaut, Time-course gene set analysis for longitudinal gene expression data, PLoS Comput. Biol. 11 (6) (2015).
[17] J. Jacques, C. Preda, Functional data clustering: a survey, Adv. Data Anal. Classif. 8 (3) (2014) 231–255.
[18] J. Jacques, C. Preda, Model-based clustering for multivariate functional data, Comput. Statist. Data Anal. 71 (2014) 92–106.
[19] A.K. Jain, Data clustering: 50 years beyond K-means, Int. Conf. Pattern Recognit. 31 (8) (2010) 651–666.
[20] G.M. James, C.A. Sugar, Clustering for sparsely sampled functional data, J. Amer. Statist. Assoc. 98 (462) (2003) 397–408.
[21] G. Kerr, H.J. Ruskin, M. Crane, P. Doolan, Techniques for clustering gene expression data, Comput. Biol. Med. 38 (3) (2008) 283–293.
[22] S. Li, N. Rouphael, S. Duraisingham, S. Romero-Steiner, S. Presnell, C.W. Davis, D.S. Schmidt, S.E. Johnson, Molecular signatures of antibody responses derived from a systems biology study of five human vaccines, Nat. Immunol. 15 (2) (2014) 195–204.
[23] N. Ling, P. Vieu, Nonparametric modelling for functional data: selected survey and tracks for future, Statistics 52 (4) (2018) 934–949.
[24] L. Liu, L. Lin, Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data, Comput. Statist. Data Anal. 138 (2019) 239–259.
[25] X. Liu, L. Wang, H. Liang, Estimation and variable selection for semiparametric additive partial linear models, Statist. Sinica 21 (3) (2011) 1225–1248.
[26] P. Ma, C.I. Castillo-Davis, W. Zhong, J.S. Liu, A data-driven clustering method for time course gene expression data, Nucleic Acids Res. 34 (4) (2006) 1261–1269.
[27] S. Ma, J. Huang, A concave pairwise fusion approach to subgroup analysis, J. Amer. Statist. Assoc. 112 (517) (2017) 410–423.
[28] S. Mankad, G. Michailidis, Biclustering three-dimensional data arrays with plaid models, J. Comput. Graph. Statist. 23 (4) (2014) 943–965.
[29] R. Opgen-Rhein, K. Strimmer, Inferring gene dependency networks from genomic longitudinal data: a functional data approach, REVSTAT 4 (2006) 53–65.
[30] J. Peng, H.-G. Müller, Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions, Ann. Appl. Stat. 2 (3) (2008) 1056–1077.
[31] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D.L. Wild, F. Falciani, Modeling T-cell activation using gene expression profiling and state-space models, Bioinformatics 20 (9) (2004) 1361–1372.
[32] D. Ruppert, Selecting the number of knots for penalized splines, J. Comput. Graph. Statist. 11 (4) (2002) 735–757.
[33] A. Schmutz, J. Jacques, C. Bouveyron, L. Cheze, P. Martin, Clustering multivariate functional data in group-specific functional subspaces, Comput. Statist. (2020) 1–31.
[34] L.L. Schumaker, Spline Functions: Basic Theory, Wiley, New York, 2007.
[35] Y.B. Slimen, S. Allio, J. Jacques, Model-based co-clustering for functional data, Neurocomputing 291 (2018) 97–108.
[36] C.J. Stone, The dimensionality reduction principle for generalized additive models, Ann. Statist. 14 (2) (1986) 590–606.
[37] A.J. Suarez, G. Subhashis, BayesIan clustering of functional data using local features, Bayesian Anal. 11 (1) (2016) 71–98.
[38] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (3) (2001) 475–494.
[39] A.W. Van Der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer New York, 1996.
[40] J.-L. Wang, J.-M. Chiou, H.-G. Müller, Functional data analysis, Annu. Rev. Stat. Appl. 3 (1) (2016) 257–295.
[41] L. Wang, L. Yang, Spline-backfitted kernel smoothing of nonlinear additive autoregression model, Ann. Statist. 35 (6) (2007) 2474–2503.
[42] H.J. Wang, Z. Zhu, J. Zhou, Quantile regression in partially linear varying coefficient models, Ann. Statist. 37 (6B) (2009) 3841–3866.
[43] J. Weiner, D.J.M. Lewis, J. Maertzdorf, H.-J. Mollenkopf, Characterization of potential biomarkers of reactogenicity of licensed antiviral vaccines: randomized controlled clinical trials conducted by the biovacsafe consortium, Sci. Rep. 9 (1) (2019) 20362.
[44] C. Wu, S. Kwon, X. Shen, W. Pan, A new algorithm and theory for penalized regression-based clustering, J. Mach. Learn. Res. 17 (1) (2015) 6479–6503.
[45] J. Xie, A. Ma, A. Fennell, Q. Ma, J. Zhao, It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data, Brief. Bioinform. 20 (4) (2019) 1450–1465.
[46] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[47] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2) (2010) 894–942.
[48] X. Zhu, A. Qu, Cluster analysis of longitudinal profiles with subgroups, Electron. J. Stat. 12 (2018) 171–193.