Journal of Multivariate Analysis 168 (2018) 221–239

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Time-varying correlation structure estimation and local-feature detection for spatio-temporal data

Xueying Zheng [a], Lan Xue [b], Annie Qu [c,d,*]

[a] Department of Biostatistics and Key Laboratory of Public Health Safety, School of Public Health, Fudan University, Shanghai 200433, China
[b] Department of Statistics, Oregon State University, Corvallis, OR 97331, USA
[c] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
[d] Department of Statistics, Fudan University, Shanghai 200433, China

ARTICLE INFO

Article history:
Received 6 October 2017
Available online 30 July 2018

Keywords:
fMRI
Local feature
Longitudinal data
Penalty
Varying-coefficient model

AMS subject classification:
62H20

ABSTRACT

Spatial–temporal data arise frequently in biomedical, environmental, political and social science studies. Capturing dynamic changes of time-varying correlation structure is scientifically important in spatio-temporal data analysis. We approximate the time-varying empirical estimator of the spatial correlation matrix by groups of selected basis matrices representing substructures of the correlation matrix. After projecting the correlation structure matrix onto a space spanned by basis matrices, we also incorporate varying-coefficient model selection and estimation for signals associated with relevant basis matrices. The unique feature of the proposed method is that signals at local regions corresponding with time can be identified through the proposed penalized objective function. Theoretically, we show model selection consistency and the oracle property in detecting local signals for the varying-coefficient estimators. The proposed method is illustrated through simulation studies and brain fMRI data.

1. Introduction

Modeling covariance structure is important for detecting associations among genes, spatial locations, social networks and brain connectivities. Developing sound spatio-temporal modeling and estimation is critical for capturing dynamic changes of associations. However, it is difficult to build a model for correlation structure for incorporating dynamic association changes that is flexible enough to capture time-varying structures, yet not burdened by high-dimensional parameter estimation. In addition, modeling spatial and temporal variations simultaneously tends to be more challenging than modeling each of them separately. Further, it is theoretically and computationally challenging to provide statistical inference for detecting dynamic changes in correlation structure.

This paper is motivated by fMRI data arising from research on children's attention deficit hyperactivity disorder (ADHD). We are interested in identifying associations and changes of associations over time among responses of brain activities from different regions in the brain. In particular, correlation structures corresponding to the regions of interest (ROIs) in the brain can change over time, although the process may be stationary, or nearly stationary. To extract the underlying signal changes of association over time, we propose a time-varying correlation structure model where dynamic changes of associations are modeled as a varying-coefficient model.

* Corresponding author at: Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.
E-mail address: anniequ@illinois.edu (A. Qu).

https://doi.org/10.1016/j.jmva.2018.07.012
0047-259X/

The related literature on local signal detection includes the fused LASSO in Tibshirani et al. [32], and the adaptive bandwidth selection approach in Miller and Hall [24] under the framework of local polynomial regression. However, as the dimension of variables increases, the aforementioned methods have computational limitations and are not effective for detecting local patterns. In addition, a functional linear regression model is proposed by James et al. [13] and Zhou et al. [42] to detect zero sub-regions. In contrast to functional data and kernel approaches, we propose a group penalized spline approach for spatio-temporal models which is able to detect dynamic changes of spatial correlations over time. The employed spline approach also has computational advantages compared to kernel approaches.

In the earlier development of spatial–temporal data analysis, most of the literature assumes that the correlation structure is fully or partially symmetric and separable, which simplifies the structure of the model. These include [6,7,17,18,27,29,30]. Recent developments on covariance estimation for spatial–temporal data also incorporate nonstationary covariance models through differential operators in environmental and disease surveillance data applications [14,19], and a generic approach to building asymmetric spatial covariance structures as in Li and Zhang [20]. In these approaches, the spatial correlation within each subject and time-varying temporal correlation information are utilized in order to make more precise statistical inference. A common feature of these approaches is that they are likelihood-based methods. However, a prior assumption on the parametric distribution might not be valid in practice. In contrast, the proposed method does not have such a restriction as it is based on the estimating equation approach [21,26] which does not require a likelihood function.

One important feature of our method is that it allows local-signal detection and coefficient estimation simultaneously for spatio-temporal data. Specifically, our approach is built on a time-varying linear representation of the inverse of the correlation matrix, projected on the span of groups of matrices basis. By introducing a group-wise penalty on varying-coefficient models, we can identify local time regions where the dynamic changes of associations occur. We show that the equivalent oracle property holds for our approach in the sense that the estimated non-zero coefficients of the basic matrices for correlation matrices can be selected consistently. The simulation studies and real data application in fMRI also confirm our theoretical findings.

The rest of the paper is organized as follows. In Section 2, the proposed method for estimating the dynamic changes of correlation structure for spatial–temporal data is presented. Section 3 establishes the asymptotic properties for the proposed estimator. Implementation strategies are illustrated in Section 4. In Section 5, we illustrate simulation studies to demonstrate the performance of the proposed method. An fMRI data set for ADHD patients is analyzed in Section 6. Finally, we provide discussion in Section 7 and proofs in the Appendix.

2. Methodology

In this section, we illustrate the method development for detecting dynamic changes using time-varying correlation matrix models for spatio-temporal data. Specifically, we transform the problem of estimating local features of the correlation structure by projecting the inverse of the correlation matrix onto the linear space spanned by groups of basis matrices and implementing piece-wise group penalization in the framework of generalized estimating equations (GEE). Through the non-parametric spline model for the coefficients of these basis matrices, the proposed model can capture dynamic changes of the correlation structure over time.

2.1. Background

For spatio-temporal data, we consider n subjects measured r_n times in the temporal dimension. At time t_{ij}, we observe an $m_i \times 1$ response vector $\mathbf{y}_i(t_{ij})$ and a $p \times m_i$ matrix $\mathbf{x}_i(t_{ij})$ of covariates for the ith subject, where $i \in \{1, \ldots, n\}, j \in \{1, \ldots, r_n\}$, and m_i indicates the number of locations or regions of interest (ROIs). To simplify notation, we assume the data is balanced with $m_i = m$, and subjects are measured at the same sequence of time points with $\{t_{ij} = t_j : j \in \{1, \ldots, r_n\}\}$. Write $\mathbf{y}_i(t_{ij})$ and $\mathbf{x}_i(t_{ij})$ as $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$, respectively.

Let $\boldsymbol{\mu}_{ij} = \mathrm{E}(\mathbf{y}_{ij}|\mathbf{x}_{ij}) = (\mu(\mathbf{x}_{ij,1}^\top \boldsymbol{\beta}), \ldots, \mu(\mathbf{x}_{ij,m}^\top \boldsymbol{\beta}))^\top$ represent the marginal mean model, where μ is a known inverse link function, $\mathbf{x}_{ij,l}$ is the p-dimensional covariate observed at ℓth ROI for ith subject at time t_j and $\boldsymbol{\beta}$ is a p-dimensional parameter vector. We generalize quasi-likelihood equations [34] for estimating the mean parameter $\boldsymbol{\beta}$ under the framework of spatio-temporal data by solving

$$\sum_{i=1}^{n} \sum_{j=1}^{r_n} \dot{\boldsymbol{\mu}}_{ij}^\top \boldsymbol{\Sigma}_{ij}^{-1} (\mathbf{y}_{ij} - \boldsymbol{\mu}_{ij}) = 0,$$

where $\dot{\boldsymbol{\mu}}_{ij} = \partial \boldsymbol{\mu}_{ij} / \partial \boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{ij} = \mathrm{cov}(\mathbf{y}_{ij})$. Liang and Zeger [21] developed the generalized estimating equation (GEE) method assuming the covariance of the response vector $\boldsymbol{\Sigma}_{ij} = \mathbf{A}_{ij}^{1/2} \mathbf{R}(t_j) \mathbf{A}_{ij}^{1/2}$, where $\mathbf{A}_{ij}$ is a diagonal matrix with the marginal variances of $\mathbf{y}_{ij}$ as the diagonal elements and $\mathbf{R}$ as a working correlation matrix. In practice, the GEE approach is robust in the sense that no specification of the full likelihood function is required.

However, a prior knowledge of the working correlation matrix is typically unavailable in practice, and the misspecified correlation can influence the efficiency of mean estimation. To achieve a more efficient estimation, Qu et al. [26] proposed the quadratic inference function (QIF), assuming that $\mathbf{R}^{-1}$ can be approximated by a linear combination of basis matrices. Zhou and Qu [40] modified the linear representation by grouping basis matrices into $\mathbf{I}_m, \mathbf{M}_1, \ldots, \mathbf{M}_D$, where $\mathbf{I}_m$ is the identity

matrix and $\mathbf{M}_1, \ldots, \mathbf{M}_D$ are subgroups of symmetric basis matrices, corresponding to different patterns of correlation structures. Examples of basis matrices are included in Section 4.1. In practice, different types of basis matrices are considered so that the combinations of these basis matrices are rich enough to approximate the inverse correlation matrix.

The rationale behind approximating the inverse of a correlation matrix by a linear combination of basis matrices is that any inverse correlation matrix can be modeled as a linear combination of a number of known basis matrices with either 0 or 1 as components. In the worst case, without knowing any correlation structure, the number of the basis matrices is equivalent to the unknown number of parameters involved in the inverse correlation matrix. In particular, Qu et al. [26] showed that common correlation matrices (e.g., exchangeable, AR1, or block correlation matrix) can be represented using a small number of basis matrices. This type of linear combination can also be useful in representing mixtures of several correlation structures [40].

In the analysis of spatio-temporal data, it is natural to assume that the correlation structure can change over time. This provides more meaningful interpretation in analyzing environmental or social networks, or fMRI data. We propose a varying-coefficient model to capture the dynamic changes of the correlation structure over time

$$\mathbf{R}^{-1}(t) \approx \boldsymbol{\alpha}_0(t)\mathbf{M}_0 + \cdots + \boldsymbol{\alpha}_D(t)\mathbf{M}_D, \tag{1}$$

where $\mathbf{M}_0 = \mathbf{I}_m$ with m being the number of regions of interest. Then for $d \in \{0, \ldots, D\}$, $\mathbf{M}_d = (\mathbf{M}_{d1}, \ldots, \mathbf{M}_{dK_d})$ consists of a group of basis matrices with different correlation structures. The number of basis matrices K_d can be different for different groups, where $K_0 = 1$. In (1), let $\boldsymbol{\alpha}_d(t) = (\alpha_{d1}(t), \ldots, \alpha_{dK_d}(t))$ represent a group of unknown coefficient functions associated with these basis matrices. In (1),

$$\boldsymbol{\alpha}_d(t)\mathbf{M}_d = \sum_{k=1}^{K_d} \alpha_{dk}(t)\mathbf{M}_{dk}$$

represents linear combinations of basis matrices for each subgroup. Although the linear combination of basis matrices is not always positive definite, the asymptotic consistency property in Section 3 ensures that the estimated correlation matrix is positive definite with high probability when the sample size is sufficiently large, which is also confirmed in our simulation studies.

The varying-coefficient model is flexible and powerful for modeling the dynamic changes of regression coefficients and has been extensively studied; see [3,8,12,25,35,37] and references therein. The coefficient functions in (1) capture the dynamic change of correlation structure over time. To estimate the unknown coefficient functions, our method applies polynomial splines to approximate them. Specifically, the spline approach approximates each nonparametric coefficient function through a linear combination of the spline bases which behave like pseudo-design variables [28]. The success of the polynomial spline relies on its good approximation power for smooth functions. Suppose the coefficient functions in (1) can be approximated, for all $k \in \{1, \ldots, K_d\}$ and $d \in \{0, \ldots, D\}$, by

$$\alpha_{dk}(t) \approx g_{dk}(t) = \sum_{h=1}^{J_n} \gamma_{dk,h} B^h(t) = \mathbf{B}^\top(t)\boldsymbol{\gamma}_{dk},$$

where $\mathbf{B}(t) = (B^1(t), \ldots, B^{J_n}(t))^\top$ with $\{B^h(t) : h \in \{1, \ldots, J_n\}\}$ being the B-spline basis and $\boldsymbol{\gamma}_{dk} = (\gamma_{dk,1}, \ldots, \gamma_{dk,J_n})^\top$. We use polynomial splines of degree p_{vc} and N_n interior knots $v_n = \{v_1, \ldots, v_{N_n}\}$, and therefore the dimension of the spline space is denoted as $J_n = N_n + p_{vc} + 1$. Then the inverse of the correlation matrix can be approximated by

$$\mathbf{R}^{-1}(t) \approx \sum_{d=0}^{D}\sum_{k=1}^{K_d} g_{dk}(t)\mathbf{M}_{d,k} = \sum_{d=0}^{D}\sum_{k=1}^{K_d} \mathbf{B}^\top(t)\boldsymbol{\gamma}_{dk}\mathbf{M}_{d,k}. \tag{2}$$

We propose to use the spline approximation since it often provides good approximation of smooth functions. More importantly, it is computationally faster and more efficient compared to the kernel-type method. However, the quality of spline approximation depends on the selection of knot sequence v_n. In this paper, we have used either equally spaced knots or a set of knots equally spaced in percentile ranks for the sake of simplicity. Although it works reasonably well in our numerical examples, in practice one can select the knot sequence more adaptively using data driven methods such as the BIC provided in Section 4.3.

2.2. Local feature selection

The coefficient functions associated with the relevant basis matrices in (2) are time-dependent and could be non-zero only for part of the time regions, and it is important to understand such dynamic change of these coefficient functions in different time regions. For example, in the fMRI study, associations among ROIs of the brain are likely to be active only for certain time regions, but not through the entire experiment. Capturing the dynamic changes of the correlation structure is equivalent to performing variable selection to detect local features instead of a global feature of the coefficient functions.

Intuitively, an efficient correlation matrix estimator should be close to the empirical correlation matrix. We propose to minimize the discrepancy in the framework of estimating equations, i.e., the difference between estimating equations using

the basis matrices representation of $\mathbf{R}^{-1}$ and the empirical inverse correlation matrix. Specifically, for the ith subject at time t_j, the discrepancy between the two sets of estimating equations with a consistent estimate $\widetilde{\boldsymbol{\beta}}$ and an empirical estimate $\widetilde{\mathbf{R}}$ is defined as

$$\mathbf{S}_{ij} = \dot{\boldsymbol{\mu}}_{ij}^{\top}(\widetilde{\boldsymbol{\beta}})\mathbf{A}_{ij}^{-1/2}\left\{\widetilde{\mathbf{R}}^{-1}(t_j) - \sum_{d=0}^{D}\sum_{k=1}^{K_d}\mathbf{B}^{\top}(t_j)\boldsymbol{\gamma}_{dk}\mathbf{M}_{dk}\right\}\mathbf{A}_{ij}^{-1/2}\{\mathbf{y}_{ij} - \boldsymbol{\mu}_{ij}(\widetilde{\boldsymbol{\beta}})\}.$$

Here the empirical correlation matrix $\widetilde{\mathbf{R}}(t_j)$ is calculated from the residuals $\mathbf{y}_{ij} - \boldsymbol{\mu}_{ij}(\widetilde{\boldsymbol{\beta}})$. We assume that individuals are observed at the same sequence of time points. Therefore the empirical correlation matrix at each time point is estimated using n residual vectors. Our method can also be used when individuals are measured at irregular and possibly subject-specific time points. Then the empirical correlation matrix is estimated using only those subjects that have measurements at these time points. However, our asymptotic theory in Section 3 requires that the number of observations at each distinct time are of the same order. Consequently, we define

$$\mathbf{U}_{ij} = \dot{\boldsymbol{\mu}}_{ij}^{\top}(\widetilde{\boldsymbol{\beta}})\mathbf{A}_{ij}^{-1/2}\widetilde{\mathbf{R}}^{-1}(t_j)\mathbf{A}_{ij}^{-1/2}\{\mathbf{y}_{ij} - \boldsymbol{\mu}_{ij}(\widetilde{\boldsymbol{\beta}})\}, \quad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, r_n\},$$

$$\mathbf{V}_{ij,dk} = \dot{\boldsymbol{\mu}}_{ij}^{\top}(\widetilde{\boldsymbol{\beta}})\mathbf{A}_{ij}^{-1/2}\mathbf{M}_{dk}\mathbf{A}_{ij}^{-1/2}\{\mathbf{y}_{ij} - \boldsymbol{\mu}_{ij}(\widetilde{\boldsymbol{\beta}})\}, \quad d \in \{0, \ldots, D\},\ k \in \{1, \ldots, K_d\}.$$

To achieve local sparsity for coefficients of selected basis matrices, we propose a piecewise penalized loss function

$$\sum_{i=1}^{n}\sum_{j=1}^{r_n}\left\|\mathbf{U}_{ij} - \sum_{d=0}^{D}\sum_{k=1}^{K_d}\mathbf{B}^{\top}(t_j)\boldsymbol{\gamma}_{dk}\mathbf{V}_{ij,dk}\right\|^2 + nr_n\sum_{d=1}^{D}\sum_{q=1}^{N_n+1}p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|), \tag{3}$$

where $\|\cdot\|$ denotes the vector ℓ_2-norm and p_{λ_n} is a penalty function with tuning parameter λ_n. Due to local properties of the B-spline basis, the $K_d \times (p_{vc}+1)$ vector $\boldsymbol{\theta}_{dq} = (\gamma_{d1,q}, \ldots, \gamma_{d1,q+p_{vc}}, \ldots, \gamma_{dK_d,q}, \ldots, \gamma_{dK_d,q+p_{vc}})$ determines the spline functions $\{g_{d1}, \ldots, g_{dK_d}\}$ on interval (v_{q-1}, v_q). Here v_q is the qth interior knot in the B-spline approximation. For $d \in \{1, \ldots, D\}$, the spline functions $\{g_{dk}(t) : k \in \{1, \ldots, K_d\}\}$ for group d are all zero on (v_{q-1}, v_q) if and only if all the elements in $\boldsymbol{\theta}_{dq}$ are shrunk to zero simultaneously. Therefore we penalize the coefficients associated with each local interval (v_{q-1}, v_q) in a group-wise fashion. Let $\widetilde{\boldsymbol{\gamma}} = \{\widetilde{\boldsymbol{\gamma}}_{dk} : k \in \{1, \ldots, K_d\}, d \in \{0, \ldots, D\}\}$ be a minimizer of the objective function (3). Then the resulting coefficient estimator with local sparsity is defined as $\widetilde{\alpha}_{dk}(t) = \mathbf{B}^{\top}(t)\widetilde{\boldsymbol{\gamma}}_{dk}$. The estimated coefficient functions can be completely zero on some local time intervals, thus the estimated correlation structure can be different at different time intervals. Therefore, the proposed method can be used to capture time-varying correlation structure for spatio-temporal data.

The loss function in (3) is capable to detect local non-zero-signal regions in the varying-coefficient model, rather than for the entire region. However, the proposed penalized loss function includes overlapping parameters across different group-penalty functions, and therefore imposes computational challenges and theoretical derivations. In addition, there are many choices of penalty function including the LASSO [31], adaptive LASSO [43], SCAD [4] or MCP [39]. Here we consider the non-convex SCAD penalty, which is defined by its first derivative

$$p'_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|) = \lambda_n\left\{\mathbf{1}(\|\boldsymbol{\theta}_{dq}\| \le \lambda_n) + \frac{(a_n\lambda_n - \|\boldsymbol{\theta}_{dq}\|)_+}{(a_n-1)\lambda_n}\mathbf{1}(\|\boldsymbol{\theta}_{dq}\| > \lambda_n)\right\}.$$

The SCAD penalty has been widely used because of its desirable properties such as unbiasedness, sparsity and continuity.

3. Asymptotic property

In this section, we investigate the rate of convergence for the varying-coefficient estimator $\widetilde{\alpha}_{dk}(t)$ by minimizing (3). Theorem 1 establishes the consistency of the proposed estimator, and Theorem 2 shows that, with probability approaching to 1 , the proposed estimator can be correctly identified as zero in the non-signal time regions. We assume that the observations from different individuals are independent of each other. For each $i \in \{1, \ldots, n\}$, observations $(\mathbf{Y}_{i1}, \mathbf{X}_{i1}), \ldots, (\mathbf{Y}_{ir_n}, \mathbf{X}_{ir_n})$ from the ith individual are viewed as a realization from a random process $\{\mathbf{Y}(t), \mathbf{X}(t) : t \in I\}$ at discrete times $t_1, \ldots, t_{r_n}$, where $\mathbf{Y}(t)$ is a vector of length m and contains measurements at m locations for time t. In the following, we assume that the number of locations m is fixed, while r_n, the number of measurements over time, is allowed to increase with the sample size n. In addition, we require the following conditions to establish the asymptotic theory.

(C1) Let $\Sigma_Y(t) = \mathrm{cov}\{\mathbf{Y}(t)\,|\mathbf{X}(t)\}$ be the conditional variance matrix of $\mathbf{Y}(t)$ given $\mathbf{X}(t)$. We assume that the eigenvalues of $\Sigma_Y(t)$ are bounded away from zero and infinity uniformly for $t \in I$. In addition, we assume that $\sup_{t\in I}\mathrm{E}\,\|\mathbf{Y}(t)\|^c < \infty$, for some sufficiently large $c > 0$.

(C2) The observation times $t_1, \ldots, t_{r_n}$ are independent with density $f_T(t)$ on I, and $f_T(t)$ is absolutely continuous and bounded away from zero and infinity uniformly over $t \in I$.

(C3) There exists a positive constant c such that $\sup_{t\in I}\|\mathbf{X}(t)\| \le c$.

(C4) There exist functions $\{\alpha_{dk}(t) : d \in \{0, \ldots, D\}, k \in \{1, \ldots, K_d\}\}$ and basis matrices $\{\mathbf{M}_{dk} : d \in \{0, \ldots, D\}, k \in \{1, \ldots, K_d\}\}$ such that, for all $t \in I$,

$$\mathbf{R}^{-1}(t) = \sum_{d=0}^{D} \sum_{k=1}^{K_d} \alpha_{dk}(t)\, \mathbf{M}_{d,k}.$$

(C5) For each $d \in \{0, \ldots, D\}$ and $k \in \{1, \ldots, K_d\}$, the coefficient function $\alpha_{dk}(t)$ is $(p_{vc}+1)$st continuously differentiable on I.

(C6) For $d \in \{1, \ldots, D\}$, let $E_d \subset I$ be the null region such that $\boldsymbol{\alpha}_d(t) = \left(\boldsymbol{\alpha}_{d1}(t), \ldots, \boldsymbol{\alpha}_{dK_d}(t)\right) = \mathbf{0}$ for all $t \in E_d$ and $\boldsymbol{\alpha}_d(t) \neq \mathbf{0}$ if $t \in (E_d)^{\complement}$. If $E_d \neq \emptyset$, we assume that $E_d = [e_{d1}, e_{d2}]$ is a closed interval. Let $\dot{\alpha}_{dk}(t)$ be the first order derivative of $\alpha_{dk}(t)$. We assume there exists a constant c such that $|\dot{\alpha}_{dk}(t)| \geq c$ for any $t \in [e_{d1} - \varepsilon, e_{d1}] \cup [e_{d2}, e_{d2} + \varepsilon]$ and a small constant $\varepsilon > 0$.

(C7) The number of knots $N_n \to \infty$ and the tuning parameters $a_n, \lambda_n \to 0$ as $n \to \infty$. In addition,

$$\frac{N_n}{nr_n} + \frac{N_n}{n} \to 0, \quad \frac{\ln(nr_n)}{nr_n N_n \lambda_n^2} \to 0, \quad \lambda_n N_n^2 \to \infty, \quad \frac{N_n^{3/2} a_n \lambda_n^2}{\rho_n} \to 0, \quad \frac{a_n \lambda_n N_n}{nr_n} \to 0,$$

where $\rho_n = \sqrt{N_n/n} + 1/\sqrt{N_n}$.

(C8) For the empirical inverse correlation matrix, we assume the eigenvalues of $\sqrt{n}\,\{\widetilde{\mathbf{R}}^{-1}(t) - \mathbf{R}^{-1}(t)\}$ are bounded for any $t \in [0, 1]$ with probability approaching 1.

Conditions (C1)–(C3) are commonly used in B-spline approaches to ensure consistency for the spline estimation of the varying-coefficient model. Similar conditions are also assumed in Huang et al. [11,12]. Condition (C5) provides the degree of smoothness on the time-varying coefficients $\alpha(\cdot)$, and Condition (C6) is needed to separate time regions between zero coefficients and nonzero coefficients. Condition (C7) assumes the convergence rates associated with the number of knots and the tuning parameters. Condition (C4) assumes that the inverse of the true correlation matrix can be well-represented by a linear combination of basis matrices.

We focus on cases where the inverse of the true correlation matrix can be well-represented by a set of basis matrices, since we are particularly interested in approximating the inverse of the true correlation matrix by basis matrices with easy-to-interpret structures. However, this assumption could be unreasonable when the inverse of the correlation cannot be well-approximated by a finite number of basis matrices, in particular in the scenario when $m \to \infty$. Therefore Condition (C8) also controls the consistency of the empirical estimator of the correlation matrix. As a special case of Theorem 1 in Lam and Fan [16], condition (C8) is satisfied when the distribution satisfies the sub-Gaussian tail conditions.

The following theorems state the asymptotic properties of the proposed penalized spline estimator of the varying coefficient functions and the oracle estimator, where the oracle estimator $\widetilde{\boldsymbol{\gamma}}^{(o)} = \{\widetilde{\gamma}_d^{(o)} : d \in \{0, \ldots, D\}\}$ is constructed by assuming all zero coefficients in $\{\gamma_{dk,h} : d \in \{0, \ldots, D\}, k \in \{1, \ldots, K_d\}, h \in \{1, \ldots, J_n\}\}$ as known to be zero and estimating the rest of the non-zero coefficients by minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{r_n} \left\| \mathbf{U}_{ij} - \sum_{d=0}^{D} \sum_{k=1}^{K_d} \sum_{h \in \mathbb{J}_{dk}} \mathbf{V}_{ij,dk} B^h(t_j) \gamma_{dk,h} \right\|^2, \tag{4}$$

where $\mathbb{J}_{dk}$ denotes the set of non-zero coefficients in γ_{dk}. The resulting oracle estimator of the coefficient functions $\alpha_{dk}(t)$ is denoted by $\widetilde{\alpha}_{dk}^{(o)}(t)$.

Theorem 1. *Under conditions (C1)–(C8), for any $d \in \{0, \ldots, D\}$, $k \in \{1, \ldots, K_d\}$, the oracle estimators $\widetilde{\alpha}_{dk}^{(o)}(t)$ satisfy*

$$\sup_{t \in I} |\widetilde{\alpha}_{dk}^{(o)}(t) - \alpha_{dk}(t)| = O_p\left(\sqrt{N_n/n + 1/N_n}\right).$$

Theorem 2. *Under conditions (C1)–(C8), for any $d \in \{0, \ldots, D\}$, $k \in \{1, \ldots, K_d\}$, there exists a local minimizer $\widetilde{\alpha}_{dk}(t)$ of (3) satisfying $\|\widetilde{\alpha}_{dk}(t) - \alpha_{dk}(t)\|_2 = O_p(\sqrt{N_n/n + 1/N_n})$, where $\|\cdot\|_2$ is defined in Eq. (6) in the Appendix. Let $\hat{E}_d = \{t \in I, \widetilde{\boldsymbol{\alpha}}_d(t) = \mathbf{0}\}$ be the corresponding null region of $\widetilde{\boldsymbol{\alpha}}_d(t)$, then $\Pr(\hat{E}_d \setminus E_d) \to 0$ as $n \to \infty$.*

In Theorem 1, we establish the asymptotic convergence rate for the oracle estimator. In addition, Theorem 2 states the existence and the convergence rate of the proposed penalized varying-coefficient estimator, and provides the asymptotic theory of model selection consistency for identifying null regions. Specifically, the proposed method ensures that null regions can be identified correctly with high probability when the sample size is sufficiently large.

4. Implementation

4.1. Examples of basis matrices

The selection of basis matrices plays a critical role in the proposed method. In this section, we provide two examples of candidate basis matrices for illustration. The first example uses a linear combination of some common correlation structures,

such as first-order auto-regressive (AR(1)) and exchangeable (EX) correlation. These correlation structures are useful to extract prior information from the empirical correlation structure which resembles these common structures. The detailed construction of these basis matrices is provided in the next section on simulation studies. The second example is motivated by Hu et al. [10], which employs the spectral representation of the inverse correlation matrix, viz.

$$\mathbf{R}^{-1}(t) \approx \alpha_0(t)\mathbf{I}_m + \sum_{d=1}^{D} \boldsymbol{\alpha}_d(t)\mathbf{e}_d(t)\mathbf{e}_d^{\top}(t),$$

where $\mathbf{e}_d(t)$ is the dth eigenvector associated with the dth largest eigenvalue of the sample correlation matrix at time t. Consequently, only the first few eigenvector-formed basis matrices $\mathbf{e}_d\mathbf{e}_d^{\top}$ are utilized to avoid parameter redundancy.

In practice, the correlation structure can be as simple as one of the examples above, but can also be a combination of the pre-determined basis matrices and eigenvector basis matrices. In addition, we can incorporate block-wise correlation structures, which decomposes the correlation matrix by several block matrices. Nevertheless, most of these strategies are determined by the information of the sample correlation matrix. We will illustrate these strategies for selecting basis matrices in the following simulation and real data examples.

4.2. An algorithm

Let $\{\widehat{\boldsymbol{\gamma}}_{dk} : d \in \{0, \ldots, D\}, k \in \{1, \ldots, K_d\}\}$ be the unpenalized estimator which minimizes the sum of squares in (3) without the penalty term. To numerically solve the penalized objective function in (3), we have used the unpenalized estimator as an initial value whenever the least squares estimation is feasible.

Let $\widehat{\boldsymbol{\theta}}_{dq}$ be the corresponding spline coefficients on interval (v_{q-1}, v_q) as defined in (3) . If the estimated $\widehat{\boldsymbol{\theta}}_{dq}$ is close to zero with $\|\widehat{\boldsymbol{\theta}}_{dq}\| < \varepsilon$ for a small number ε, then we set $\widehat{\boldsymbol{\theta}}_{dq}$ as zeros in the next step. In the following, we describe a quadratic approximation to solve (3) for the non-zero coefficients. A quadratic approximation [4] of p_{λ_n} at non-zero $\boldsymbol{\theta}^*_{dq}$ in (3) is defined as

$$p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|) \approx p_{\lambda_n}(\|\boldsymbol{\theta}^*_{dq}\|) + \frac{1}{2}\, p'_{\lambda_n}(\|\boldsymbol{\theta}^*_{dq}\|)\, \|\boldsymbol{\theta}^*_{dq}\|^{-1}\{\boldsymbol{\theta}_{dq}^{\top}\boldsymbol{\theta}_{dq} - (\boldsymbol{\theta}^*_{dq})^{\top}\boldsymbol{\theta}^*_{dq}\},$$

where p' is the first derivative of the penalty function. Then the loss function (3) can be approximated (up to a constant) by

$$\sum_{i=1}^{n}\sum_{j=1}^{r_n}\left\|\mathbf{U}_{ij} - \sum_{d=0}^{D}\sum_{k=1}^{K_d}\mathbf{V}_{ij,dk}\mathbf{B}^{\top}\boldsymbol{\gamma}_{dk}\right\|^2 + \frac{nr_n}{2}\sum_{d=0}^{D}\sum_{q=1}^{N_n+1} c_{dq}\,\boldsymbol{\theta}_{dq}^{\top}\boldsymbol{\theta}_{dq},$$

where $c_{dq} = p'_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|)\|\boldsymbol{\theta}_{dq}\|^{-1}$. Consequently, the nonzero components of $\boldsymbol{\gamma}_{dk}$ can be updated by minimizing the above quadratic function. We denote

$$C_{dq} = \sum_{k=j-p_{vc}}^{q} c_{dq}, \quad \mathbf{C}_d = (C_{d1}, \ldots, C_{dJ_n})^{\top}, \quad \mathbf{C} = (\mathbf{C}_1^{\top}, \ldots, \mathbf{C}_D^{\top})^{\top},$$

where $c_{dk} = 0$ for $k < 1$ or $k > N_n + 1$. Let $\hat{\boldsymbol{\gamma}}^{k+1}$ be the solution at the $(k+1)$st iteration. We update the non-zero components in $\hat{\boldsymbol{\gamma}}^{k+1}$ by

$$\hat{\boldsymbol{\gamma}}^{k+1} = \{(\mathbf{V}_s^{*,k})^{\top}\mathbf{V}_s^{*,k} + nr_n\mathrm{diag}(\mathbf{C}_s)\}^{-1}(\mathbf{V}_s^{*,k})^{\top}\mathbf{U}_s^k,$$

where $\mathbf{V}_s^{*,k}$ contains the columns of $\mathbf{V}_n^{*,k}$ corresponding to the nonzero components of $\hat{\boldsymbol{\gamma}}^k$, and $\mathbf{C}_s$ and $\mathbf{U}_s^k$ are similarly defined, while $\mathbf{V}_n^{*,k}$ and $\mathbf{U}_n^k$ are defined in the Appendix, but evaluated at $\hat{\boldsymbol{\gamma}}^k$.

The proposed algorithm consists of two steps. In the first step, we select basis matrices by minimizing the Euclidean norm of the discrepancy between two sets of estimating equations using the empirical inverse correlation matrix and the linear representation of basis matrices with a grouped SCAD penalty. In the second step, we employ a local-feature selector (3) to detect zero-subregion coefficients in the varying-coefficient model to refine the local correlation structure estimation.

The two-step strategy can be simply merged into one step if the number of parameters is not large, which minimizes the loss function (3) directly based on the entire space spanned by all candidate basis matrices. This strategy leads to significant drawbacks compared with the two-step method. The major drawbacks are that the correlation structure is not clearly represented through specific basis matrices, which leads to significant computational burden. In general, the proposed two-step approach is computationally more feasible, and effective for capturing the dynamic change of the correlation structure.

4.3. Tuning parameters selection

To implement the proposed method, we need to choose the knots to control the smoothness of the spline-estimated curve and λ_n to determine the complexity of the selected model. Choice of these tuning parameters has a crucial effect on the performance of the proposed group-penalized varying-coefficient model. For simplicity, we use equally spaced knots, and choose the number of interior knots N_n to be the integer part of $(nmr_n)^{1/(2p_{vc}+3)}$, where p_{vc} is the degree of polynomial

spline. One can also select the number of interior knots N_n using a data-driven procedure, such as the BIC described below. Here we only focus on the selection of λ_n in the penalty function for computation simplicity. Similar strategies have been applied by Huang et al. [12], He et al. [9] and Xue and Qu [37].

We use the Bayesian Information Criteria (BIC) procedure to select λ_n, and fix $a_n = 3.7$ as suggested by Fan and Li [4]. Following Qu and Li [25], we denote the estimator of the coefficients γ as $\hat{\gamma}_{\lambda_n}$. Let

$$\hat{\mathbf{U}}_{ij}(\lambda_n) = \sum_{d=0}^{D}\sum_{k=1}^{K_d}\mathbf{V}_{ij,dk}\mathbf{B}^{\top}\hat{\boldsymbol{\gamma}}_{dk}(\lambda_n)$$

and z_n be the total number of nonzero components in $\hat{\boldsymbol{\gamma}}_{\lambda_n}$. Then we choose $\hat{\lambda}_n$ that minimizes the BIC value

$$\hat{\lambda}_n = \text{argmin}_{\lambda_n}\text{BIC}(\lambda_n) = \text{argmin}_{\lambda_n}\left[\ln\left[\frac{1}{nr_n}\sum_{i=1}^{n}\sum_{j=1}^{r_n}\left\{\mathbf{U}_{ij} - \hat{\mathbf{U}}_{ij}(\lambda_n)\right\}^2\right] + \frac{\ln(nr_n)z_n}{nr_n}\right].$$

5. Simulation

In this section we conduct simulation studies to illustrate the performance of the proposed local-feature selection method. When generating correlated spatio-temporal data, we allow spatial correlation structure to vary over time with different magnitudes of correlation for both continuous and binary responses. The performance of the proposed method under various settings of cluster (and subject) size is investigated and compared with the penalized varying-coefficient model without incorporating local features proposed in Xue and Qu [37]. In addition, we also simulate unbalanced clustered data with some missing observations to confirm that the proposed method can still perform well for unbalanced data. In the last simulation study, we examine the proposed method when the basis matrices are intentionally misspecified.

5.1. Study 1: normal responses

We generate spatio-temporal data from a regression model with a normally distributed error term by setting, for all $i \in \{1, \ldots, 100\}$ and $j \in \{1, \ldots, 34\}$,

$$\mathbf{y}_{ij} = \beta_0 + \beta_1\mathbf{x}_{ij,1} + \beta_2\mathbf{x}_{ij,2} + \beta_3\mathbf{x}_{ij,3} + \boldsymbol{\varepsilon}_{ij},$$

where the observation times $t_1, \ldots, t_{34}$ are uniform within the interval $[0, 1]$; $\mathbf{y}_{ij}$ is the response vector with measurements taken at $m = 25$ or 75 spatial locations for the ith subject at time t_j; $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (2, 1, 1, 1)$; the covariates $\{\mathbf{x}_{ij,k}; k \in \{1, 2, 3\}\}$ are $m \times 1$ vectors with elements generated independently from the standard Normal distribution; and $\boldsymbol{\varepsilon}_{ij}$ is the $m \times 1$ error term following a multivariate normal distribution $\mathcal{N}[0, \mathbf{R}(t_j)]$.

The spatial correlation matrix $\mathbf{R}(t_j)$ is assumed to have a diagonal structure of the form such that $\mathbf{R}(t_j) = \text{diag}\{\mathbf{R}_1(t_j), \ldots, \mathbf{R}_5(t_j)\}$, where each diagonal block $\mathbf{R}_1(t_j), \ldots, \mathbf{R}_5(t_j)$ is of dimension 5×5 or 15×15 corresponding to cluster size $m = 25$ or 75 respectively. Furthermore, the first and second blocks, $\mathbf{R}_1(t_j)$ and $\mathbf{R}_2(t_j)$, have AR(1) and exchangeable (EX) correlation structure, respectively. The remaining three blocks have an independent structure with $\{\mathbf{R}_k(t_j) : k \in \{3, 4, 5\}\}$ being identity matrices. In the following, we allow the magnitude of correlation in both $\mathbf{R}_1(t)$ and $\mathbf{R}_2(t)$ to be time-varying.

We specify eleven basis matrices of the block-diagonal matrices with sub-block size of 5 when $m = 25$, or sub-block size of 15 when $m = 75$. For example, when $m = 25$, the first group $\{\mathbf{M}_0, \ldots, \mathbf{M}_4\}$ contains the identity matrix $\mathbf{M}_0 = \mathbf{I}_{25}$ (25×25 identity matrix) and four matrices with $\mathbf{I}_5$ (5×5 identity matrix) on the first to the fourth diagonal blocks respectively, and zero entries elsewhere. The second group $\{\mathbf{M}_5, \mathbf{M}_6\}$ contains two matrices to represent the AR(1) structure for the first block and zero for the rest of the blocks, where $\mathbf{M}_5$ has 1s on the sub-diagonal in the first block and 0's elsewhere, and $\mathbf{M}_6$ has 1s on two corner components of the diagonal in the first block and 0s elsewhere. The third group includes only one matrix $\mathbf{M}_7$ for the first block and 0s elsewhere, corresponding to the exchangeable correlation structure for the first block with 1s on the off-diagonal and 0s elsewhere. The fourth and fifth groups ($\{\mathbf{M}_8, \mathbf{M}_9\}$ and $\{\mathbf{M}_{10}\}$) are defined similarly as the second and third groups respectively, but defined for the second block instead.

To generate time-varying correlation structure, we notice that $\mathbf{R}^{-1}(t_j) = \text{diag}\{\mathbf{R}_1^{-1}(t_j), \ldots, \mathbf{R}_5^{-1}(t_j)\}$, which can be exactly presented as linear combinations of basis matrices in the first, second and fifth groups defined above. In particular, the coefficients of basis matrices in the second and fifth groups are of the form such that

$$\alpha_5(t) = \begin{cases}(-4t^2 + 4t - 1)\rho_{\text{AR}} & \text{if } t > 0.5,\\ 0 & \text{if } t < 0.5,\end{cases} \quad \alpha_{10}(t) = \begin{cases}(-4t^2 + 4t - 1)\rho_{\text{EX}} & \text{if } t < 0.5,\\ 0 & \text{if } t > 0.5,\end{cases} \tag{5}$$

where larger values of ρ_{AR} and ρ_{EX} are associated with higher correlations in the corresponding block. We generate weak, medium and strong levels of correlation with parameters $(\rho_{\text{AR}}, \rho_{\text{EX}}) = (0.50, 0.25), (1.50, 0.70)$, or $(2, 1)$, which correspond to maximum correlations of 0.4, 0.6 and 0.8, respectively. In this example, the true correlation structure is different at time intervals $(0, 0.5)$ and $(0.5, 1)$. Specifically, the first half-time region is of AR(1) correlation structure in the first block and independent correlation structure in the other blocks, and the magnitude of correlation monotonically decreases

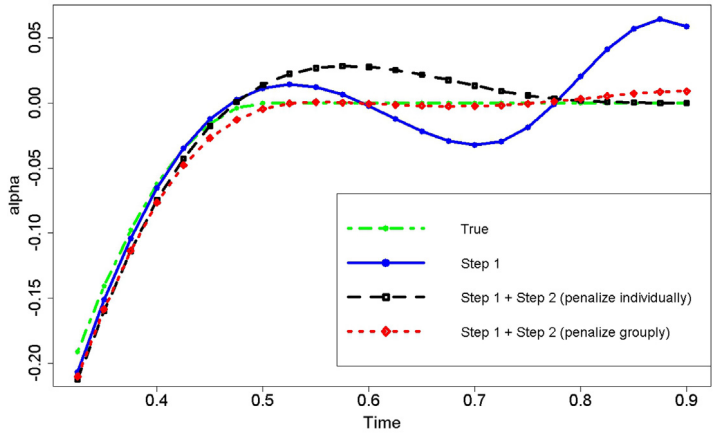

Fig. 1. The global and local estimators for α_5. The blue line is the global penalized estimator without incorporating local features. The black line indicates an individual parameter penalization after global penalization. The red line is the proposed estimator which penalizes by groups of parameters after global penalization. The green dotted line is the true value of α_5.

Table 1
The percentages of correct-identification (C), over-identification (O) and under-identification (U) for varying-coefficients for normal responses with sample size $n = 100$ in Study 1.

Cluster no.	Scenario	α_5			α_{10}		
		C	O	U	C	O	U
		Local					
	(II)Weak	0.848	0.119	0.033	0.855	0.083	0.062
	(III)Medium	0.778	0.215	0.007	0.826	0.159	0.016
	(IV)Strong	0.764	0.234	0.002	0.823	0.173	0.004
m = 25		Global					
	(II)Weak	0.500	0.000	0.000	0.502	0.001	0.497
	(III)Medium	0.511	0.029	0.460	0.541	0.454	0.005
	(IV)Strong	0.513	0.486	0.001	0.532	0.465	0.003
		Local					
	(II)Weak	0.745	0.218	0.037	0.865	0.023	0.112
	(III)Medium	0.733	0.258	0.009	0.882	0.052	0.066
	(IV)Strong	0.697	0.301	0.002	0.905	0.078	0.017
m = 75		Global					
	(II)Weak	0.500	0.000	0.500	0.504	0.016	0.480
	(III)Medium	0.500	0.000	0.500	0.593	0.286	0.121
	(IV)Strong	0.504	0.491	0.005	0.585	0.401	0.014

to zero when time increases from 0 to 0.5. In the second half-region, the correlation structure in the second block is exchangeable with monotonically increasing correlation as time increases, while the other blocks all have independent correlation structure.

We apply the proposed method to identify dynamic spatial correlation structures over time. Specifically, we select four knots equally spaced in the interval [0, 1] and adopt the quadratic spline. Note that the non-parametric spline model with a higher degree of polynomial function can intensify the degree of overlapping parameters in the group penalty function, and might be more computationally time-consuming. The simulations are repeated 200 times for each set-up.

We first compare the proposed local-feature selection method to the existing global model selection method proposed in Xue and Qu [37] for varying coefficient models. Without incorporating local features, the global method shrinks the coefficient functions to be completely zero in the entire time region, which is equivalent to implementing only Step 1 in the proposed algorithm. Fig. 1 plots the estimates of $\alpha_5(t)$ using these two methods. It shows that the proposed local-feature selection method estimates the coefficient function reasonably well. In particular, the estimated zero sub-region using the local-feature selection method almost coincides with the true zero sub-region. In contrast, the global method without incorporating local features performs poorly in identifying the zero region, which is consistent with the results reported in Table 1. In addition, another possible approach for local-feature selection is to penalize the spline coefficients for each individual basis matrix separately in (3). The resulting estimator is plotted as the black dotted line in Fig. 1. It indicates that after the first-step selection of basis matrices, penalizing the associated coefficients individually is less efficient for capturing the exact local signal compared to the proposed group penalization of coefficients corresponding to the group basis matrices method.

Table 2
The percentages of correct-identification (C), over-identification (O) and under-identification (U) local feature selection method for binary responses with cluster size $m = 25$ in Study 2.

Sample size	Scenario	α_5			α_{10}		
		C	O	U	C	O	U
		Balanced					
	(II)Weak	0.824	0.169	0.007	0.816	0.125	0.059
	(III)Medium	0.819	0.176	0.005	0.812	0.146	0.042
	(IV)Strong	0.795	0.203	0.002	0.806	0.158	0.036
$n = 100$		Unbalanced, 15% missing					
	(II)Weak	0.792	0.121	0.027	0.718	0.113	0.169
	(III)Medium	0.754	0.241	0.005	0.783	0.168	0.049
	(IV)Strong	0.747	0.250	0.003	0.783	0.185	0.032
		Balanced					
	(II)Weak	0.897	0.070	0.033	0.734	0.034	0.232
	(III)Medium	0.866	0.127	0.007	0.884	0.076	0.040
	(IV)Strong	0.854	0.144	0.002	0.875	0.106	0.019
$n = 200$		Unbalanced, 35% missing					
	(II)Weak	0.701	0.293	0.006	0.732	0.203	0.065
	(III)Medium	0.702	0.294	0.005	0.713	0.243	0.044
	(IV)Strong	0.702	0.296	0.002	0.706	0.273	0.021

The percentages of correct-identification (C), over-identification (O) and under-identification (U) for signal regions of the coefficient functions are presented in Table 1. We compare the proposed method (local) with the existing penalized varying-coefficient model selection without incorporating local features (global). If the absolute value of the coefficient estimator is less than 0.01 for the selected zero region or is greater than 0.05 for the non-zero region, then we count them as correct identification. Here over-identification (O) is defined as true zero coefficients estimated as non-zero signals, and under-identification (U) is defined as true non-zero coefficients estimated as zeros. We observe that our approach performs consistently well since the percentages of correct-identification (C) are mostly above 70% in all scenarios. In addition, the over-identification percentage increases when the signal is stronger. This is because the sharp increase of the correlation signal in the interval immediately after $t = 0.5$ imposes a great challenge to local feature detection. A possible solution to obtain more accurate local estimation around the change point is to add additional knots. Table 1 also indicates that the proposed method performs well when cluster size $m = 75$, even though the large cluster size adds difficulty in capturing the time-varying signal.

In addition, the varying-coefficient model incorporating local features performs better in detecting local features than the global selection method, as shown in Table 1. For example, the percentage of correct-identification ranges between 75% to 90% when $m = 25$ for the proposed local-feature approach, in contrast to only 50% correct-identification for the global varying-coefficient model approach in weak, medium and strong scenarios. The improvement of correct-identification is more than 50% through utilizing the local-feature penalty.

5.2. Study 2: Binary responses with unbalanced data

In this study, we consider a binary response for spatio-temporal data. The binary response is generated from the logistic regression model defined, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, 34\}$, by

$$\text{logit}(\boldsymbol{\mu}_{ij}) = \beta_0 + \mathbf{x}_{ij,1}^{\top}\beta_1 + x_{ij,2}^{\top}\beta_2 + x_{ij,3}^{\top}\beta_3,$$

where $\boldsymbol{\mu}_{ij} = \text{E}(\mathbf{y}_{ij}|\mathbf{x}_{ij})$ and $\mathbf{y}_{ij}$ is the binary response vector measured at 25 locations for the ith subject at time t_j, the covariates $\mathbf{x}_{ij} = \{x_{ij,k} : k \in \{1, 2, 3\}\}$ with each covariate being a 1×25 vector with elements independently generated from $\mathcal{N}(0, 0.01)$, and the regression coefficients $\boldsymbol{\beta} = (1, 0.05, 0.05, 0.05)$. The R package `mvtBinaryEP` is used to generate the binary response vector $\mathbf{y}_{ij}$ using the above logistics regression model with the same time-varying correlation structures as the ones for the normal case. The same pre-specified basis matrices are also used in the estimation procedure.

The proposed method can be implemented to accommodate an unbalanced spatio-temporal data, where the missing mechanism is completely random. The estimation procedure can adopt the idea of the transformation matrix to each cluster; see more details in Zhou and Qu [40]. We conduct numerical studies to illustrate the feasibility of the proposed method to handle unbalanced spatio-temporal data. The unbalanced data are generated by randomly deleting observations from a balanced spatio-temporal data. For each balanced data set, 30% (when $n = 100$) or 70% (when $n = 200$) of the entire clusters of $m = 25$ measurements are randomly selected. For these selected clusters, 50% of the observations within each cluster are deleted completely at random. Consequently, 15% of the observations are missing when $n = 100$ and 35% of the observations are missing when $n = 200$, respectively.

The percentages of correct-identification (C), over-identification (O) and under-identification (U) for signal regions of coefficient functions are presented in Table 2, which indicates that the proposed method detects the signal region effectively for both balanced and unbalanced binary responses. When we increase the sample size to $n = 200$, the proposed method

Table 3
The percentages of correct-identification (C), over-identification (O) and under-identification (U) local feature selection method for complex correlation structure with cluster size $m = 25$ in Study 3.

Coefficient	Weak			Strong		
	C	O	U	C	O	U
	n = 100					
α_5	0.764	0.229	0.007	0.760	0.233	0.007
α_{10}	0.845	0.137	0.018	0.836	0.145	0.019
α_{11}	0.762	0.071	0.166	0.861	0.095	0.044
	n = 200					
α_5	0.836	0.155	0.009	0.832	0.158	0.009
α_{10}	0.910	0.071	0.019	0.909	0.072	0.020
α_{11}	0.861	0.025	0.114	0.932	0.036	0.032

Table 4
The loss of estimated correlation matrix $\hat{\mathbf{R}}$ and empirical correlation matrix $\widetilde{\mathbf{R}}$, $\|\hat{\mathbf{R}} - \mathbf{R}\|_F$ and $\|\widetilde{\mathbf{R}} - \mathbf{R}\|_F$ in Frobenius norm in Study 3.

Sample size	Weak		Strong	
	Loss	sd	Loss	sd
		$\hat{R}$ (local)		
100	0.290	0.549	0.288	0.414
200	0.075	0.044	0.090	0.071
		$\widetilde{R}$ (empirical)		
100	6.029	0.087	6.018	0.088
200	3.009	0.045	3.003	0.045

performs better than when $n = 100$, which supports the consistency of local feature selection when the sample size is sufficiently large. Since a substantial part of the observations are missing for the unbalanced data sets, the proposed method performs slight less effectively for the unbalanced data compared to the fully observed cases for $n = 100$ and 200. However, Table 2 indicates that the proposed method is still quite effective with correct-identification above 70%, even when 50% of the observations among 70% of the clusters are missing for $n = 200$.

5.3. Study 3: complex correlation structure

In this study, we evaluate the performance of the proposed method when the basis matrices are misspecified. We generate data from the same regression model as the one in Study 1 for the normal case, but with different correlation structures. Specifically, we model the correlation structure at time t as $\mathbf{R}(t) = \mathbf{R}_1(t) + \mathbf{R}_2(t)$, where $\mathbf{R}_1(t)$ is a 25×25 matrix of the same block diagonal structure as defined in study 1, and $\mathbf{R}_2(t)$ is a 25×25 matrix with $\rho_{\text{off}}(t)$ in entries $(i, i+5)$ and $(i+5, i)$ for $i \in \{1, \ldots, 5\}$, and 0 elsewhere. We consider $\mathbf{R}_1(t)$ to be at weak or strong levels of correlation with $(\rho_{\text{AR}}, \rho_{\text{EX}}) = (0.50, 0.25)$ or $(2, 1)$ in Eq. (5). The coefficient $\rho_{\text{off}}(t)$ is zero on interval $[0, 0.5]$ and linearly increases from 0 to 0.25 in the weak-signal scenario, or from 0 to 0.5 in the strong-signal scenario, on time interval $(0.5, 1)$. The matrix $\mathbf{R}_2(t)$ is designed under the constraint of positive-definite correlation matrix. Here we add one more basis matrix $\mathbf{M}_{11}$ (associated with α_{11}) to the original pool of basis matrices, with 1s in the entries $(i, i+5)$ and $(i+5, i)$ for $i \in \{1, \ldots, 5\}$, and 0 elsewhere. This leads to the case where basis matrices could be misspecified, since the off-diagonal $\mathbf{R}_2(t)$ affects the structure of $\mathbf{R}^{-1}(t)$, and the inverse correlation matrix could be unstructured.

In Table 3, we provide the percentages of correct-identification (C), over-identification (O) and under-identification (U) to illustrate the performance of local-feature model selection for the unstructured inverse correlation setting. The proposed approach performs well on feature selection for all of α_5, α_{10} and α_{11}. Even when the true correlation matrix cannot be exactly captured by a linear combination of the selected basis matrices, the local features of the coefficients associated with the basis matrices are consistently detected when the sample size increases.

Since the inverse of a hybrid correlation matrix does not have a specific structure, and the misspecification of basis matrices occurs in this study, we measure the difference between the estimated $\hat{\mathbf{R}}$ and the true $\mathbf{R}$ through the loss of

$$\frac{1}{34\,m}\sum_{i=1}^{34}\|\hat{\mathbf{R}}(t) - \mathbf{R}(t)\|_F$$

in Table 4, where $\|\mathbf{R}\|_F$ is the Frobenius norm defined by the square root of the trace of the squared matrix. For $n = 100$ and 200, Table 4 shows that the proposed correlation matrix estimators have less bias than the empirical estimators. The losses based on the empirical correlation matrix estimators are significantly larger than the losses based on the local-feature estimators.

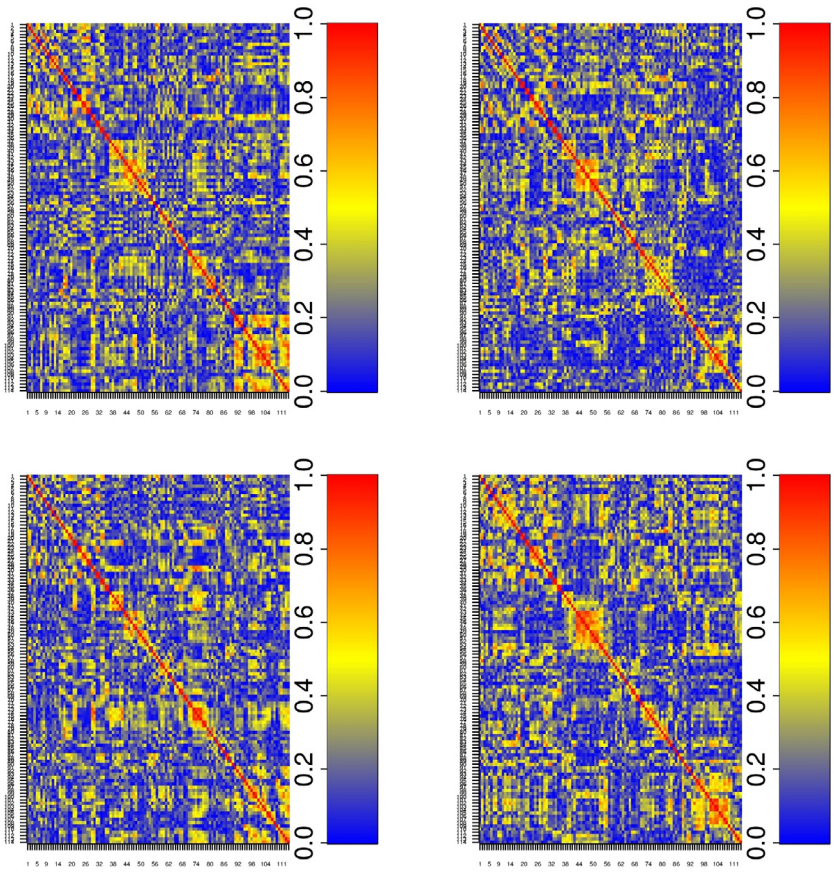

Fig. 2. Heat maps of the associations of 116 regions of interest at the 1st, 4th, 7th and 10th time points.

6. FMRI data analysis

In this section we implement our method to an attention deficit hyperactivity disorder (ADHD) data set. ADHD is one of the psychiatric disorders commonly found in children and adolescents with characteristics of being easily distracted, impulsive, and restless. The fMRI data measured in the resting state of 79 patients from ADHD-200 test samples are collected from Oregon Health and Science University (OHSU) data. Among them, 42 patients are typically-developing controls, 30 are ADHD combined, 5 are ADHD inattentive and 2 are ADHD hyperactive. We removed one subject due to missing observations, and consequently the total number of subjects is 78. The fMRI data is measured from 116 regions of interest (ROIs) of the brain over 74 time points, which are processed by the automated anatomical labeling (AAL) software package and digital atlas designed for the human brain.

For the fMRI studies, statistical challenges [22] such as dealing with massive data, balancing computational feasibility with model efficiency and explaining brain connectivity are still under-developed. Specifically, there has been an increasing demand for understanding the brain, especially for modeling and quantifying the interaction among different ROIs of the brain. Grinband et al. [5] pointed out that modeling temporal variability of the fMRI data can increase the statistical power and capture an important source of information on the relationship between brain activity and psychophysical performance. Cribben et al. [1] utilized a data-driven technique of graphical LASSO, detected temporal change points in functional connectivity, and estimated graphical relationships among ROIs under the assumption of a conditional normal distribution. Recent literature focusing on the spatio-temporal correlation modeling of the brain fMRI data also includes, but is not limited to, Lindquist [23] and Kang et al. [15].

Here we identify the dynamic temporal correlation change by the proposed local penalized varying-coefficient method. Fig. 2 illustrates the associations and connectivity of 116 ROIs at four time points (the 1st, 4th, 7th and 10th time points) using heat maps of empirical correlation matrices. Note that the red color indicates stronger associations and the blue color indicates no associations among different regions of interest. We are able to identify some patterns of associations for several blocks of the regions of interest, where the associations show dynamic change over time.

We fit a regression model with time-varying spatial correlation within a subject by setting, for each $i \in \{1, \ldots, 78\}$ and $j \in \{1, \ldots, 74\}$,

$$\mathbf{y}_{ij} = \beta_0 + \beta_1 \mathbf{x}_{ij,1} + \cdots + \beta_5 \mathbf{x}_{ij,5} + \varepsilon_{ij},$$

where $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij,k}$ for $k \in \{1, \ldots, 5\}$ are the response and covariates observed from the ith subject at time t_{ij}. Specifically, $\mathbf{y}_{ij}$ is a response vector of length 116 measured at these ROIs, and the covariates include gender, age, diagnosis (DX), and whether patients are in the category of inattentive or impulsive. All the covariates are observed at the subject level and do not vary with time and ROIs. The focus of our analysis is to utilize the proposed local-feature selection method to construct a time-varying approximation for the spatial correlation structure.

From the heat maps in Fig. 2, we observe that the correlation matrices are not sparse, and a standard correlation model might not be able to capture the high variability of brain function. For better illustration, in Fig. 3, we plot the inverse correlation matrices at 9 time points selected from the original 74 time points representing the whole period of scanning, specifically at the 8th, 11th, 14th, 20 th, 23rd, 28th, 56th, 72nd and 74th time points. We also choose a sub-region of 23 RoIs, including 6, 5, 4 and 8 RoIs from the frontal, occipital, parietal and vermis regions of the brain, respectively. The connectivity information can be imputed from the heat map of the empirical estimator of the inverse correlation matrices in Fig. 3. Apparently, we observe that the signals along the diagonal blocks of the matrix are generally stronger than the off-diagonal signals. In fact, most of the strong signals can be divided into several blocks on the diagonal, corresponding to those basis matrices of certain blocks.

As a result, we split 23 RoIs into 4 blocks with sizes 6, 5, 4 and 8 respectively. These consist of 14 basis matrices $\{\mathbf{M}_0, \ldots, \mathbf{M}_{13}\}$, including 4 basis matrices of identity and AR(1) structure for each block, 4 eigenvector-based-block matrices and 2 eigenvector-based matrices for the entire matrix. The second basis matrix for the AR(1) structure (with 1s on the two corner components) does not play an important role for estimation, and therefore is omitted here. We choose 2 equally spaced knots and adopt the cubic spline method. As a result, the total number of parameters (γ) involved is $14 \times 6 = 84$. We select the tuning parameter as 0.142 for the SCAD penalty.

We plot the heat map constructed from the proposed estimator of the inverse correlation matrices in Fig. 4. The heat map indicates that there are several block structures including diagonal and off-diagonal blocks for the inverse correlation matrix, while the empirical inverse correlation matrix estimator only shows a higher correlation along the diagonal blocks. The findings of the proposed estimator are consistent with that of the real experiment. That is, the variation of brain connectivity function at the beginning and ending stages is relatively stronger than during the middle stage in the experiment. This phenomenon is more obvious with the proposed method than the empirical estimator. In addition, we notice that the within-block dynamic change of correlation structure is not always in accordance with the entire regions of interest' correlation structure. Specifically, the associations among the 5 RoIs in the occipital area are more intense during the most time of the experiment, and the time-dependent within-block signals from the frontal (the first block) and parietal (the third block) regions also display strong connections, while the off-diagonal spatial connections are relatively weak in the middle period of the experiment. In summary, the proposed local feature selection method can effectively detect signal and non-signal regions of dynamic correlation structure using a relatively small number of basis matrices.

7. Discussion

The time-varying correlation structure model is flexible and powerful for identifying time-dependent associations for spatio-temporal data. In this paper, we develop a local penalized varying-coefficient model to effectively quantify and detect dynamic changes from the spatial correlation structure. One distinct feature of the proposed approach is that we are able to incorporate local features of a nonparametric function, and provide local-signal detection and estimation simultaneously for spatio-temporal data. Our simulation studies and data application to fMRI data indicate that the penalized nonparametric varying-coefficient model for the inverse of correlation matrix can capture dynamic changes of associations within several groups of regions of interest and identify time intervals when the dynamic changes of spatial associations occur simultaneously.

To produce local sparse coefficient estimators, our proposed method approximates the coefficient functions using spline functions and penalizes the ℓ_2 norm of spline coefficients associated with each subinterval spanned by the knot sequences. As pointed out by a referee, other types of penalization may also be considered, including direct Tikhonov penalization of the higher-order derivatives of the covariance functions. In addition, the proposed group penalization process involving overlapping parameters is solved by the quadratic approximation algorithm. However, it is also critical to develop more computationally efficient algorithms to model spatio-temporal data when the number of spatial locations increases, since high-dimensional matrix operations are computationally costly. Furthermore, theoretical derivation for local feature model selection of the correlation structure is quite challenging when the cluster size diverges [33]. These topics are worth further investigation in future research.

Acknowledgments

The authors thank the Editor-in-Chief, an Associate Editor, and two reviewers for insightful comments, and suggestions which improve the article significantly. Xueying Zheng 's research is supported by the National Natural Science Foundation of China (No. 11501124). Annie Qu 's research is supported by US National Science Foundation (DMS-1308227, DMS-1415308 and DMS-1613190).

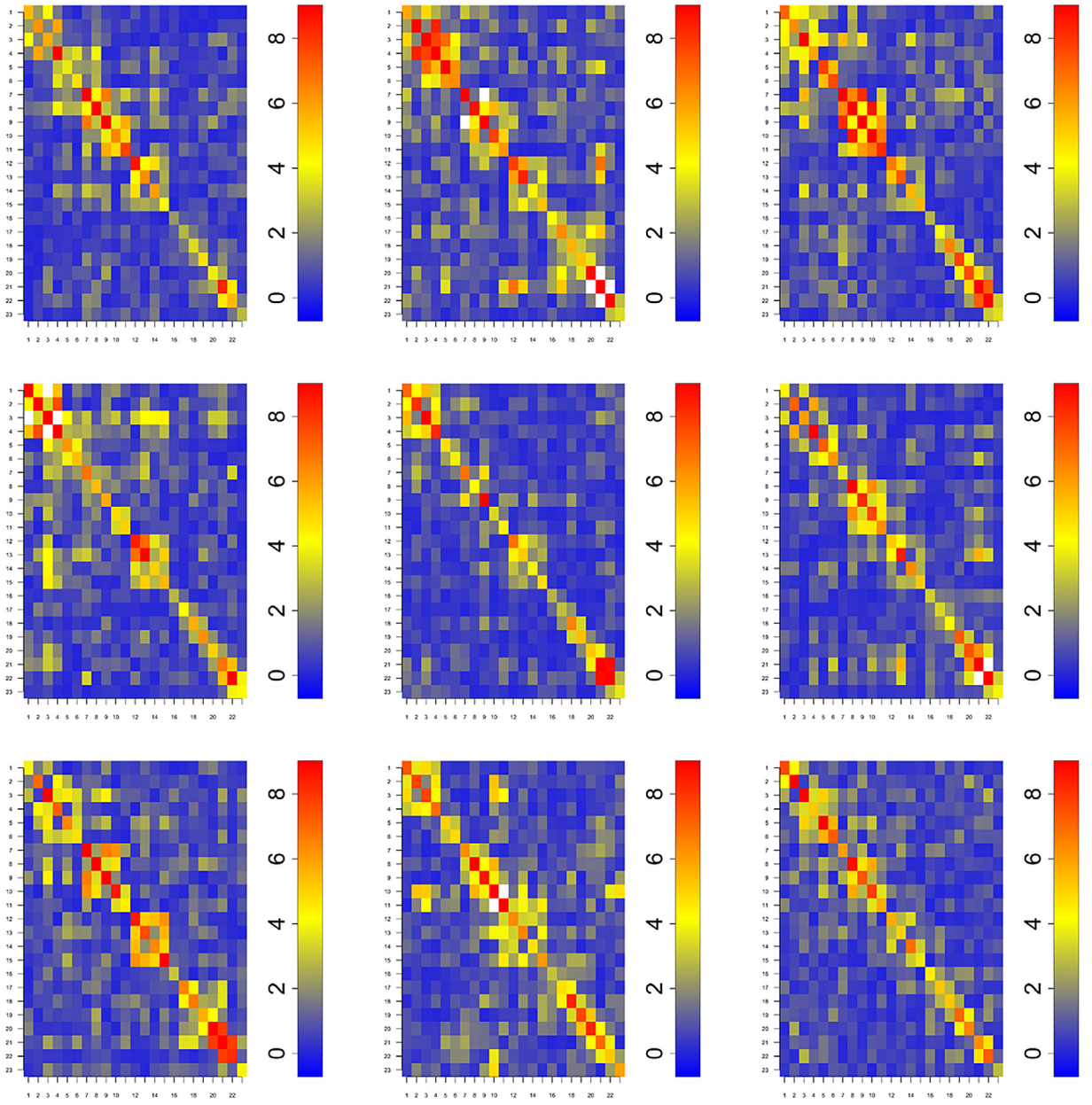

Fig. 3. Heat maps of 23×23 empirical estimator of time-varying inverse correlation matrices at nine selected time points.

Appendix

A.1. Notation and lemmas

Let $C^{p_{vc}}$ be the space of p_{vc} times continuously differentiable functions on $[0, 1]$, and G_n be the spline approximation space of order p_{vc} with interior knots $v_n = \{v_1, \ldots, v_{N_n}\}$. We denote any positive constants by the same notations of c and C (with $c < C$) without distinction in each case.

Let $\mathcal{M}$ be the model space as a collection of $D_J = K_0 + \cdots + K_D$ vectors of functions

$$\mathcal{M} = \{\boldsymbol{\alpha}(t) = \{\alpha_{dk}(t), d \in \{0, \ldots, D\},\ k \in \{1, \ldots, K_d\} : \alpha_{dk}(t) \in C^{p_{vc}+1}\}\},$$

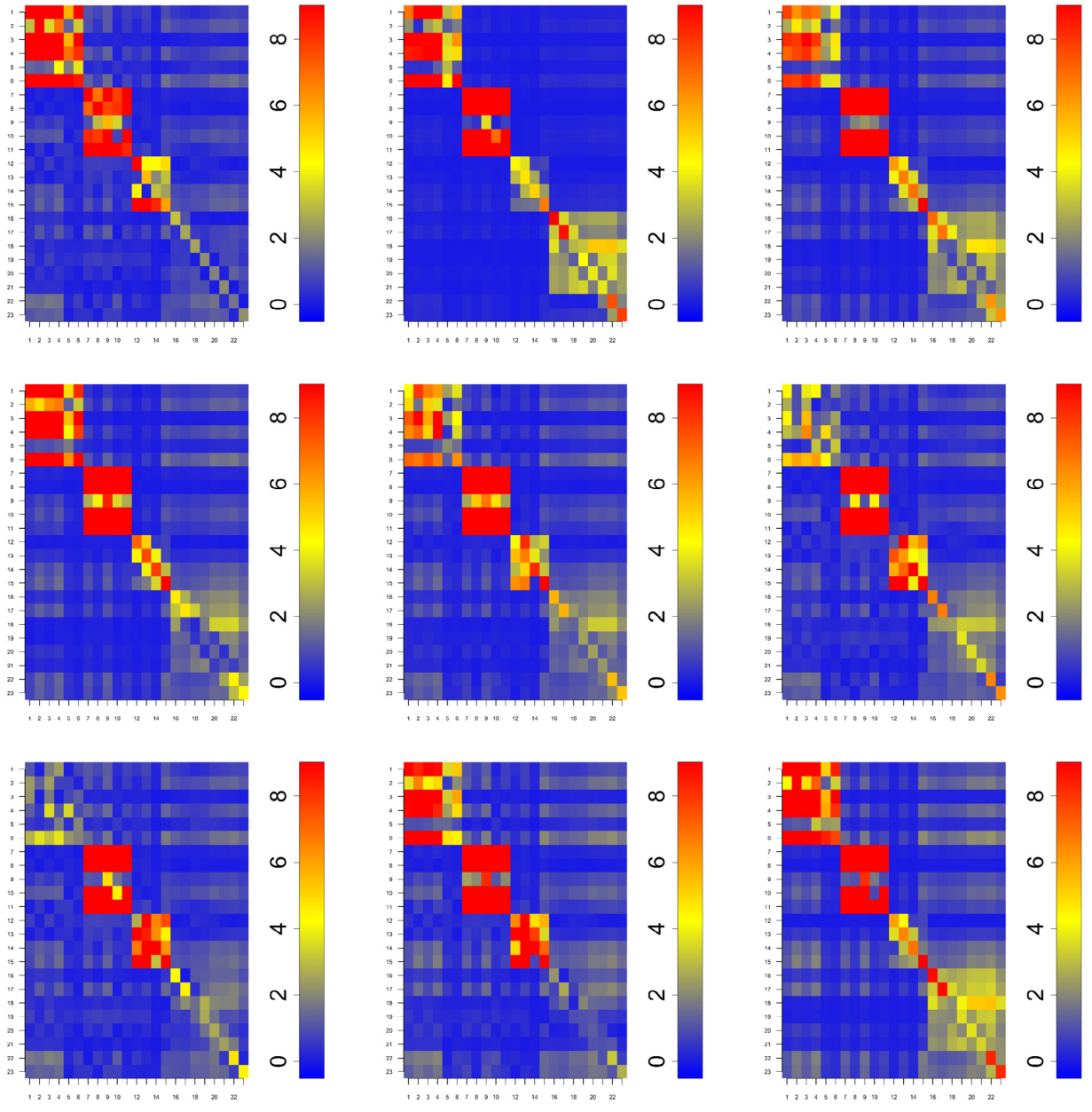

Fig. 4. Heat maps of 23 × 23 proposed estimator of time-varying inverse correlation matrices at nine selected time points.

and let the approximation space be defined similarly as

$$\mathcal{M}_n = \{g(t) = \{g_{dk}(t), d \in \{0, \ldots, D\}, k \in \{1, \ldots, K_d\} : g_{dk}(t) \in G_n\}\}.$$

Let $\mathbf{V}^*_{ij,dk} = \mathbf{V}_{ij,dk}\mathbf{B}^\top(t_j)$ be a $p \times J_n$ matrix, $\mathbf{V}^*_{ij,d} = (\mathbf{V}^*_{ij,d1}, \ldots, \mathbf{V}^*_{ij,dK_d})$, $\mathbf{V}^*_{ij} = (\mathbf{V}^*_{ij,0}, \ldots, \mathbf{V}^*_{ij,D})$, and $\boldsymbol{\gamma}_d = (\boldsymbol{\gamma}^\top_{d1}, \ldots, \boldsymbol{\gamma}^\top_{dK_d})^\top$, for $d \in \{0, \ldots, D\}$. Let $\mathbf{U}_i = (\mathbf{U}^\top_{i1}, \ldots, \mathbf{U}^\top_{ir_n})^\top$ and $\mathbf{V}^*_i = (\mathbf{V}^{*\top}_{i1}, \ldots, \mathbf{V}^{*\top}_{ir_n})^\top$ for $i \in \{1, \ldots, n\}$. Then $\boldsymbol{\gamma}_n = (\boldsymbol{\gamma}^\top_0, \ldots, \boldsymbol{\gamma}^\top_D)^\top$ is a J_nD_J-dimensional vector, $\mathbf{U}_n = (\mathbf{U}^\top_1, \ldots, \mathbf{U}^\top_n)^\top$ is an nr_np -dimensional vector, and $\mathbf{V}^*_n = (\mathbf{V}^{*\top}_1, \ldots, \mathbf{V}^{*\top}_n)^\top$ is an $nr_np \times J_nD_J$ matrix. Then the corresponding loss function (3) is equivalent to

$$\left\|\mathbf{U}_n - \mathbf{V}^*_n\boldsymbol{\gamma}_n\right\|^2 + nr_n \sum_{d=1}^{D}\sum_{q=1}^{N_n+1} p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|).$$

For any $\alpha \in \mathcal{M}$, define the theoretical and empirical norms on $\mathcal{M}$ as

$$\|\boldsymbol{\alpha}\|_2^2 = \mathrm{E}\{\boldsymbol{\alpha}^\top(T)\boldsymbol{\alpha}(T)\} = \sum_{d=0}^{D}\sum_{k=1}^{K_d}\mathrm{E}\{\alpha_{dk}^2(T)\},$$
$$\|\boldsymbol{\alpha}\|_n^2 = \frac{1}{nr_n}\sum_{i=1}^{n}\sum_{j=1}^{r_n}\{\boldsymbol{\alpha}^\top(t_j)\boldsymbol{\alpha}(t_j)\} = \sum_{d=0}^{D}\sum_{k=1}^{K_d}\left\{\frac{1}{nr_n}\sum_{i=1}^{n}\sum_{j=1}^{r_n}\alpha_{dk}^2(t_j)\right\}, \tag{6}$$

respectively. The following Lemmas 1 and 2 are obtained similarly as Lemma A.4 of Xue and Yang [38].

Lemma 1. *Under conditions (C1)–(C5), for any $g \in \mathcal{M}_n$, there exist constants $0 < c < C$, such that $c\|\boldsymbol{\gamma}_n\|^2/N_n \le \|g\|_2^2 \le C\|\boldsymbol{\gamma}_n\|^2/N_n$.*

Lemma 2. *Under conditions (C1)–(C5), for any $g \in \mathcal{M}_n$, there exist constants $0 < c < C$, such that, as $n \to \infty$, $\Pr(c\|g\|_2^2 \le \|g\|_n^2 \le C\|g\|_2^2) \to 1$.*

Lemma 3. *Under conditions (C1)–(C5), there exist constants $0 < c < C$, such that for any vector of γ_n of length J_nD_J, we have, as $n \to \infty$,*

$$\Pr\{c\|\boldsymbol{\gamma}_n\|^2/N_n \le \boldsymbol{\gamma}_n^\top\mathbf{V}_n^{*\top}\mathbf{V}_n^*\boldsymbol{\gamma}_n/(nr_n) \le C\|\boldsymbol{\gamma}_n\|^2/N_n\} \to 1.$$

Proof. For any $\boldsymbol{\gamma}_n$ of length J_nD_J, we have

$$\frac{1}{nr_n}\boldsymbol{\gamma}_n^\top\mathbf{V}_n^{*\top}\mathbf{V}_n^*\boldsymbol{\gamma}_n = \frac{1}{nr_n}\sum_{i=1}^{n}\sum_{j=1}^{r_n}\boldsymbol{\gamma}_n^\top\mathbf{V}_{ij}^{*\top}\mathbf{V}_{ij}^*\boldsymbol{\gamma}_n = \frac{1}{nr_n}\sum_{i=1}^{n}\sum_{j=1}^{r_n}\sum_{d=0}^{D}\sum_{k=1}^{K_d}\sum_{h=1}^{J_n}\mathbf{V}_{ij,dk}^\top\mathbf{V}_{ij,dk}\{B^h(t_j)\gamma_{dk,h}\}^2.$$

Given conditions (C1) and (C3), there exist constants $0 < c_1 \le C_1$, such that

$$\frac{c_1}{nr_n}\boldsymbol{\gamma}_n^\top\mathbf{B}_n^\top\mathbf{B}_n\boldsymbol{\gamma}_n \le \frac{1}{nr_n}\boldsymbol{\gamma}_n^\top\mathbf{V}_n^{*\top}\mathbf{V}_n^*\boldsymbol{\gamma}_n \le \frac{C_1}{nr_n}\boldsymbol{\gamma}_n^\top\mathbf{B}_n^\top\mathbf{B}_n\boldsymbol{\gamma}_n,$$

where $\mathbf{B}_n = (\mathbf{B}^\top(t_1), \ldots, \mathbf{B}^\top(t_r))^\top$ is an $nr_n \times J_nD_J$ matrix of B-spline basis and $\mathbf{B}(t_j) = \mathbf{B}^\top(t_j) \otimes \mathbf{1}_{D_j}$ is a $1 \times J_nD_J$ vector. By Lemma 6.2 of Zhou et al. [41], there exist constants $0 < c_2 \le C_2$, such that

$$\frac{c_2}{N_n}\|\boldsymbol{\gamma}_n\|^2 \le \frac{1}{nr_n}\boldsymbol{\gamma}_n^\top\mathbf{B}_n^\top\mathbf{B}_n\boldsymbol{\gamma}_n \le \frac{C_2}{N_n}\|\boldsymbol{\gamma}_n\|^2,$$

with probability approaching 1 as $n \to \infty$. The Lemma follows by taking $C = C_1C_2$ and $c = c_1c_2$.. □

Let $G_{dk}^{(o)} \subset G_n$ be the oracle spline approximation space containing spline functions with zero values on the null region E_d. The following lemma can be proved by the approximation theory in de Boor [2].

Lemma 4. *Under conditions (C1)–(C5), there exists a spline function $g_{dk}^{(o)} \in G_{dk}^{(o)}$, such that*

$$\sup_{t\in(0,1)}|\alpha_{dk}(t) - g_{dk}^{(o)}(t)| = O(N_n^{-1}).$$

Proof. The approximation theory in de Boor [2] entails that there exists a spline function $g_{dk} \in G_n$ such that

$$\sup_{t\in(0,1)}|\alpha_{dk}(t) - g_{dk}(t)| = O\{N_n^{-(p_{vc}+1)}\},$$

where

$$g_{dk}(t) = \sum_{h=1}^{N_n+p_{vc}+1}\gamma_{dk,h}B^h(t)$$

for a set of coefficients $\{\gamma_{dk,h} : h \in \{1, \ldots, N_n + q + 1\}\}$. For the null region $\mathrm{E}_d = [e_{d1}, e_{d2}]$, suppose e_{d1} falls in the interval between the ℓ_{d1}th and $(\ell_{d1}+1)$st knots, and e_{d2} falls in the interval between the ℓ_{d2} th and $(\ell_{d2}+1)$st knots. Let $\mathbb{J}_d = \{1, \ldots, \ell_{d1} - 1, \ell_{d1} + p_{vc} + 2, \ldots, J_n\}$. Now let $g_{dk}^*(t) = \sum_{h\in\mathbb{J}_d}\gamma_{dk,h}B^h(t)$. Let $\widetilde{E}_d = [\ell_{d1}, \ell_{d2+1}]$, $A_d = \widetilde{E}_d \setminus E_d$. Then $g_{ij}^* \in G_{ij}^{(o)}$, and

$$\sup_{t\in E_d}|\alpha_{dk}(t) - g_{dk}^*(t)| = 0, \quad \sup_{t\in\widetilde{E}_d^c}|\alpha_{dk}(t) - g_{dk}^*(t)| = O\{N_n^{-(p_{vc}+1)}\},$$

and

$$\sup_{t\in A_d}\left|\alpha_{dk}(t)-g^*_{dk}(t)\right| \le \sup_{t\in A_d}|\alpha_{dk}-g_{dk}|+\sup_{t\in A_d}|g_{dk}-g^*_{dk}|$$
$$\le 2\sup_{t\in(0,1)}|\alpha_{dk}-g_{dk}|+\sup_{t\in A_d}|\alpha_{dk}| = O\{N_n^{-(p_{vc}+1)}+N_n^{-1}\} = O(N_n^{-1}).$$

Putting these three cases together, one has $\sup_{t\in(0,1)}|\alpha_{dk}(t)-g^{(o)}_{dk}(t)| = O(N_n^{-1})$. □

A.2. Proof of Theorem 1

To prove Theorem 1, we define $\widetilde{\boldsymbol{\gamma}}^*$ as a minimizer of (4), but with the true $\mathbf{R}^{-1}(t)$, instead of $\widetilde{\mathbf{R}}^{-1}(t)$ in (4). Then, for $t\in[0,1]$, let $\widetilde{\alpha}^*_{dk}(t)=\mathbf{B}^\top(t)\widetilde{\boldsymbol{\gamma}}^*_{dk}$, for $d\in\{0,\ldots,D\}$ and $k\in\{1,\ldots,K_d\}$.

By Lemma 4, there exist spline functions $g^{(o)}_{dk}\in G^{(o)}_{dk}$, such that $\|\alpha_{dk}-g^{(o)}_{dk}\|_\infty \le cN_n^{-1}$ for a constant c. Let

$$\mathbf{M}_d\mathbf{g}^{(o)}_d=\sum_{k=1}^{K_d}g^{(o)}_{dk}\mathbf{M}_{dk},\quad \mathbf{M}_d\boldsymbol{\alpha}_d=\sum_{k=1}^{K_d}\alpha_{dk}\mathbf{M}_{dk}$$

and

$$\mathbf{U}_{ij}(\mathbf{g}^{(o)})=\sum_{d=0}^{D}\dot{\mu}_{ij}\mathbf{A}_{ij}^{-1/2}\mathbf{M}_d g^{(o)}_d\mathbf{A}_{ij}^{-1/2}(\mathbf{y}_{ij}-\boldsymbol{\mu}_{ij}),\quad \mathbf{U}_{ij}(\boldsymbol{\alpha})=\sum_{d=0}^{D}\dot{\mu}_{ij}\mathbf{A}_{ij}^{-1/2}\mathbf{M}_d\boldsymbol{\alpha}_d\mathbf{A}_{ij}^{-1/2}(\mathbf{y}_{ij}-\boldsymbol{\mu}_{ij}),$$

where $i\in\{1,\ldots,n\}$ and $j\in\{1,\ldots,r_n\}$. Define $\boldsymbol{\epsilon}^I_{ij}=\mathbf{U}_{ij}(\widetilde{\mathbf{R}}^{-1})-\mathbf{U}_{ij}(\mathbf{R}^{-1})$, $\boldsymbol{\epsilon}^{II}_{ij}=\mathbf{U}_{ij}(\mathbf{R}^{-1})-\mathbf{U}_{ij}(\mathbf{g}^{(o)})$. Let $\boldsymbol{\epsilon}^I_i=(\boldsymbol{\epsilon}^{I,\top}_{i1},\ldots,\boldsymbol{\epsilon}^{I,\top}_{ir_n})^\top$, $\mathbf{E}^I=(\boldsymbol{\epsilon}^{I,\top}_1,\ldots,\boldsymbol{\epsilon}^{I,\top}_n)^\top$, and define $\mathbf{E}^{II}$ similarly. Then we have

$$\begin{aligned}\widetilde{\alpha}^{(o)}(t)-\alpha(t)&=\widetilde{\alpha}^{(o)}(t)-\widetilde{\alpha}^*(t)+\widetilde{\alpha}^*(t)-g^{(o)}(t)+g^{(o)}(t)-\alpha(t)\\&=\mathbf{B}^\top(t)(\mathbf{V}_n^{*\top}\mathbf{V}_n^*)^{-1}\mathbf{V}_n^{*\top}\mathbf{E}^I+\mathbf{B}^\top(t)(\mathbf{V}_n^{*\top}\mathbf{V}_n^*)^{-1}\mathbf{V}_n^{*\top}\mathbf{E}^{II}+g^{(o)}(t)-\alpha(t)\\&=I(t)+II(t)+III(t).\end{aligned}$$

For $I(t)$, by Lemma 3 and the Cauchy–Schwarz inequality, we have

$$|I(t)|\le\sqrt{\mathbf{B}^\top(t)(\mathbf{V}_n^{*\top}\mathbf{V}_n^*)^{-1}\mathbf{B}(t)}\sqrt{\left(\mathbf{E}^I\right)^\top\mathbf{V}_n^*(\mathbf{V}_n^{*\top}\mathbf{V}_n^*)^{-1}\mathbf{V}_n^{*\top}\mathbf{E}^I}\le c\frac{N_n}{nr_n}\sqrt{\mathbf{B}^\top(t)\mathbf{B}(t)}\sqrt{(\mathbf{E}^I)^\top\mathbf{V}_n^*\mathbf{V}_n^{*\top}\mathbf{E}^I}.$$

We notice $\sup_{t\in(0,1)}\sqrt{\mathbf{B}^\top(t)\mathbf{B}(t)}\le\sqrt{D_J(p_{vc}+1)}$ as the upper bound of B-spline bases is 1 and each B-spline base has support on only $p_{vc}+1$ intervals. Together with conditions (C1), (C3), (C8), we have

$$\sqrt{(\mathbf{E}^I)^\top\mathbf{V}_n^*\mathbf{V}_n^{*\top}\mathbf{E}^I}=\sqrt{\sum_{d=0}^{D}\sum_{k=1}^{K_d}\sum_{h=1}^{J_n}\left\{\sum_{i=1}^{n}\sum_{j=1}^{r_n}B^h_{ij}(\mathbf{y}_{ij}-\boldsymbol{\mu}_{ij})^\top\mathbf{A}_{ij}^{-1/2}\mathbf{M}_{d,k}\mathbf{A}_{ij}^{-1/2}\dot{\boldsymbol{\mu}}_{ij}^\top\dot{\boldsymbol{\mu}}_{ij}\mathbf{A}_{ij}^{-1/2}(\widetilde{\mathbf{R}}^{-1}-\mathbf{R}^{-1})\mathbf{A}_{ij}^{-1/2}(\mathbf{y}_{ij}-\boldsymbol{\mu}_{ij})\right\}^2}$$
$$\le c\sqrt{\frac{1}{n}}\sqrt{\sum_{d=0}^{D}\sum_{k=1}^{K_d}\sum_{h=1}^{J_n}\left(\sum_{i=1}^{n}\sum_{j=1}^{r_n}B^h_{ij}\right)^2}=O_p\left(\sqrt{\frac{1}{n}}\right)O_p\left(\frac{nr_n}{\sqrt{N_n}}\right)=O_p\left(\sqrt{\frac{n}{N_n}}r_n\right).$$

Therefore

$$\sup_{t\in(0,1)}|I(t)|=\frac{N_n}{nr_n}O_p\left(\sqrt{\frac{n}{N_n}}r_n\right)=O_p\left(\sqrt{\frac{N_n}{n}}\right).\tag{7}$$

Similarly, under the condition (C4), one has

$$\sup_{t\in(0,1)}|II(t)|=\frac{N_n}{nr_n}O_p\left(\frac{nr_n}{\sqrt{N_n}}\frac{1}{N_n}\right)=O_p\left(1/\sqrt{N_n}\right).\tag{8}$$

Further, since $\sup_{t\in(0,1)}|III(t)|=\sup_{t\in(0,1)}|g^{(o)}_{dk}(t)-\alpha_{dk}(t)|\le c(N_n^{-1})$, together with (7) and (8), we have

$$\sup_{t\in(0,1)}|\widetilde{\boldsymbol{\alpha}}^{(o)}(t)-\boldsymbol{\alpha}(t)|=\sup_{t\in(0,1)}|I(t)+II(t)+III(t)|=O_p(\sqrt{N_n/n}+1/\sqrt{N_n}),$$

which guarantees the uniform convergence theorem of the oracle estimator for the spline coefficient functions by minimizing (4), i.e.

$$\sup_{t\in(0,1)}|\widetilde{\alpha}^{(o)}_{dk}(t)-\alpha_{dk}(t)|=O_p(\sqrt{N_n/n+1/N_n}).$$

This completes the argument. □

A.3. Proof of Theorem 2: Consistency of varying coefficient estimators

Let

$$\ell(\boldsymbol{\gamma}) = \|\mathbf{U}_n - \mathbf{V}_n^*\boldsymbol{\gamma}\|^2, \quad P\ell(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}) + nr\sum_{d=1}^{D}\sum_{q=1}^{N_n+1} p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|), \quad \ell_{dk,h}(\boldsymbol{\gamma}) = \frac{\partial \ell(\boldsymbol{\gamma})}{\partial \gamma_{dk,h}}$$

and $\bar{\boldsymbol{\ell}}_{dk,h}(\boldsymbol{\gamma}) = (\ell_{dk,h}(\boldsymbol{\gamma}), \ldots, \ell_{dk,h+p_{vc}}(\boldsymbol{\gamma}))$. Let $\rho_n = \sqrt{N_n/n} + 1/\sqrt{N_n}$. For any $\boldsymbol{\gamma}$ satisfying $\|\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}^{(o)}\| = c\rho_n\sqrt{N_n}$ with some constant $c > 0$, one has

$$\begin{aligned} P\ell(\boldsymbol{\gamma}) - P\ell(\widetilde{\boldsymbol{\gamma}}^{(o)}) &= \ell(\boldsymbol{\gamma}) - \ell(\widetilde{\boldsymbol{\gamma}}^{(o)}) + nr_n\sum_{d=1}^{D}\sum_{q=1}^{N_n+1} p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|) - nr_n\sum_{d=1}^{D}\sum_{q=1}^{N_n+1} p_{\lambda_n}(\|\widetilde{\boldsymbol{\theta}}_{dq}^{(o)}\|) \\ &= \|\mathbf{U}_n - \mathbf{V}_n^*(\widetilde{\boldsymbol{\gamma}}^{(o)} + \mathbf{u})\|^2 - \|\mathbf{U}_n - \mathbf{V}_n^*\widetilde{\boldsymbol{\gamma}}^{(o)}\|^2 + nr_n\sum_{d=1}^{D}\sum_{q=1}^{N_n+1}\{p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|) - p_{\lambda_n}(\|\widetilde{\boldsymbol{\theta}}_{dq}^{(o)}\|)\} \\ &= I + II, \end{aligned}$$

in which, by the arguments as in the proof of Theorem 2 in Xue [36], there exists a constant $c > 0$, such that $I = \|\mathbf{U}_n - \mathbf{V}_n^*(\widetilde{\boldsymbol{\gamma}}^{(o)} + \mathbf{u})\|^2 - \|\mathbf{U}_n - \mathbf{V}_n^*\widetilde{\boldsymbol{\gamma}}^{(o)}\|^2 \geq cnr_n\rho_n^2$ with probability approaching 1. For the second part, let $s_{d0} = \{j : \|\widetilde{\boldsymbol{\theta}}_{dq}^{(o)}\| = 0\}$, $s_{d1} = \{j : a_n\lambda_n < \|\widetilde{\boldsymbol{\theta}}_{dq}^{(o)}\|, a_n\lambda_n < \|\boldsymbol{\theta}_{dq}\|\}$, and $s_{d2} = (s_{d0} \cup s_{d1})^{\complement}$. Then one has,

$$II \geq nr_n\sum_{d=1}^{D}\sum_{q\in s_{d2}}\{p_{\lambda_n}(\|\boldsymbol{\theta}_{dq}\|) - p_{\lambda_n}(\|\widetilde{\boldsymbol{\theta}}_{dq}^{(o)}\|)\} \approx nr_nD\,(a_n\lambda_nN_n)\,\lambda_n(\|\boldsymbol{\theta}_{dq}\| - \|\widetilde{\boldsymbol{\theta}}_{dq}^{(o)}\|) = O_p(nr_nN_n^{3/2}a_n\lambda_n^2\rho_n) = o_p(nr_n\rho_n^2),$$

by assumption (C7). Therefore, for some constant c, one has

$$\Pr\{\inf_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\widetilde{\boldsymbol{\gamma}}^{(o)}\|=c\rho_n\sqrt{N_n}} P\ell(\boldsymbol{\gamma}) \geq P\ell(\widetilde{\boldsymbol{\gamma}}^{(o)})\} \to 1, \text{ as } n \to \infty.$$

Hence there exists a local minimizer $\tilde{\boldsymbol{\gamma}}$ of the penalized objective function $P\ell(\boldsymbol{\gamma})$ in the local neighborhood $\{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}^{(o)}\| \leq c\rho_n\sqrt{N_n}\}$. Together with Theorem 1, there exists a local minimizer $\widetilde{\alpha}_{dk}(t)$ of (3) satisfying $\|\widetilde{\alpha}_{dk} - \alpha_{dk}\|_2 = O_p(\sqrt{N_n/n} + 1/\sqrt{N_n})$. □

A.4. Proof of Theorem 2: Consistency of null region identification

Note that for each null region $E_d = [e_{d1}, e_{d2}]$ defined in (C6), there exist two knots, i.e. the ℓ_{d1}th and ℓ_{d2}th knots, such that e_{d1} falls in the interval between the ℓ_{d1}th and $(\ell_{d1}+1)$st knots, and e_{d2} falls in the interval between the ℓ_{d2}th and $(\ell_{d2}+1)$st knots. Let $\mathbb{J}_d = \{1, \ldots, \ell_{d1} - 1, \ell_{d2} + p_{vc} + 2, \ldots, J_n\}$. Then we have the following Lemma by Lemma 4 and Markov's inequality.

Lemma 5. *Under conditions (C2)–(C5), for any $d \in \{0, \ldots, D\}$ and $k \in \{1, \ldots, K_d\}$, let*

$$\widetilde{\ell}_{dk,h}(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{r_n}\mathbf{V}_{ij,kh}(t_j)B^h(t_j)\left\{\mathbf{U}_{ij}(t_j) - \sum_{k=0}^{D}\sum_{d=0}^{K_d}\sum_{h\in\mathbb{J}_d}\mathbf{V}_{ij,kh}(t_j)B^h(t_j)\boldsymbol{\gamma}_{dk,h}\right\}.$$

Let $\bar{\boldsymbol{\ell}}_{dk,q}(\boldsymbol{\gamma}) = (\widetilde{\ell}_{dk,q}(\boldsymbol{\gamma}), \ldots, \widetilde{\ell}_{dk,q+p_{vc}}(\boldsymbol{\gamma}))$ and $\bar{\boldsymbol{\ell}}_{d,q}(\boldsymbol{\gamma}) = (\bar{\ell}_{d1,q}(\boldsymbol{\gamma}), \ldots, \bar{\ell}_{dd_j,q}(\boldsymbol{\gamma}))$. Then for any η_n satisfying

$$\frac{1}{\eta_n}\sqrt{\frac{\ln(nr_n)}{nrN_n}} \to 0 \quad \textit{and} \quad \eta_nN_n^2 \to \infty,$$

we have

$$\Pr\left\{\max_{d\in\{1,\ldots,D\},\, q\in\mathbb{J}_d^{\complement}}\left\|\bar{\boldsymbol{\ell}}_{d,q}(\widetilde{\boldsymbol{\gamma}}^{(o)})\right\| \geq nr_n\eta_n\right\} \to 0.$$

Let $\boldsymbol{\gamma}_{dk,\mathbb{J}_{dk}} = (\boldsymbol{\gamma}_{dk,h}, h \in \mathbb{J}_{dk})^\top$ and $\boldsymbol{\gamma}_{dk,\mathbb{J}_{dk}^{\complement}} = (\boldsymbol{\gamma}_{dk,h}, h \in \mathbb{J}_{dk}^{\complement})^\top$.

Let $\widetilde{\boldsymbol{\gamma}} = \{\widetilde{\boldsymbol{\gamma}}_{dk} : d \in \{0, \ldots, D\}, k \in \{1, \ldots, K_d\}\}$ such that, for each $\widetilde{\boldsymbol{\gamma}}_{dk}$ with $\widetilde{\boldsymbol{\gamma}}_{dk,\mathbb{J}_{dk}^{\complement}} = 0$, and $\widetilde{\boldsymbol{\gamma}}_{dk,\mathbb{J}_{dk}}$ solving $\partial P\ell(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}_{dk,h} = 0$, for $h \in \mathbb{J}_{dk}$. By the Karush–Kuhn–Tucker (KKT) condition, $\widetilde{\boldsymbol{\gamma}}$ is the minimizer of (3) with the SCAD penalty if and only if

$$\begin{cases} \dfrac{\partial P\ell(\widetilde{\boldsymbol{\gamma}})}{\partial\gamma_{dk,h}} = 0 & \text{if } \|\widetilde{\boldsymbol{\theta}}_{dq}\| \neq 0, \\ \|\bar{\ell}_d(\widetilde{\boldsymbol{\gamma}})\| \leq nr_n\lambda_n & \text{if } \|\widetilde{\boldsymbol{\theta}}_{dq}\| = 0, \end{cases}$$

where $\bar{\boldsymbol{\ell}}_d(\widetilde{\boldsymbol{\gamma}}) = (\bar{\boldsymbol{\ell}}_{d1,q}(\widetilde{\boldsymbol{\gamma}}), \ldots, \bar{\ell}_{dd_j,q}(\widetilde{\boldsymbol{\gamma}}))$ and $\bar{\boldsymbol{\ell}}_{dk,q}(\widetilde{\boldsymbol{\gamma}}) = (\ell_{dk,q}(\widetilde{\boldsymbol{\gamma}}), \ldots, \ell_{dk,q+p_{vc}}(\widetilde{\boldsymbol{\gamma}}))$. By its definition, the first condition in the KKT is satisfied by $\widetilde{\boldsymbol{\gamma}}$. One only needs to show that $\|\bar{\boldsymbol{\ell}}_{dk}(\widetilde{\boldsymbol{\gamma}})\|_2 \le \lambda_n$, if $\|\widetilde{\boldsymbol{\gamma}}_d^q\| = 0$.

Let $A = \cup_{d,k} \mathbb{J}_{dk}^{\mathbb{C}}$ and $\widetilde{\boldsymbol{\gamma}}_A = (\widetilde{\boldsymbol{\gamma}}_{dk,h}, h \in A)$. Define $\widetilde{\boldsymbol{\gamma}}_{A^{\mathbb{C}}}$, $\widetilde{\boldsymbol{\gamma}}_A^{(o)}$ and $\widetilde{\boldsymbol{\gamma}}_{A^{\mathbb{C}}}^{(o)}$ similarly. Then note that $\widetilde{\boldsymbol{\gamma}}_{A^{\mathbb{C}}} = \widetilde{\boldsymbol{\gamma}}_{A^{\mathbb{C}}}^{(o)} = 0$, and

$$\widetilde{\boldsymbol{\gamma}}_A^{(o)} = \left\{ \left(\mathbf{V}_{n,A}^*\right)^\top \mathbf{V}_{n,A}^* \right\}^{-1} \left\{ \left(\mathbf{V}_{n,A}^*\right)^\top \mathbf{U}_{n,R} \right\}, \quad \widetilde{\boldsymbol{\gamma}}_A = \left\{ \left(\mathbf{V}_{n,A}^*\right)^\top \mathbf{V}_{n,A}^* \right\}^{-1} \left\{ \left(\mathbf{V}_{n,A}^*\right)^\top \mathbf{U}_{n,R} + \mathbf{W}_n \right\},$$

where $\mathbf{W}_n$ is a diagonal matrix with $\mathbf{W}_n = \mathrm{diag}\{p'_{\lambda_n}(\|\widetilde{\boldsymbol{\theta}}_{dq}\|)\widetilde{\boldsymbol{\theta}}_{dq}/\|\widetilde{\boldsymbol{\theta}}_{dq}\|\}$. Note that

$$\|\mathbf{W}_n\|_2 = \sum_{d,q:\|\widetilde{\boldsymbol{\theta}}_{dq}\| \neq 0} p'_{\lambda_n}(\|\widetilde{\boldsymbol{\theta}}_{dq}\|) \le c\,(a_n \lambda_n N_n)\,\lambda_n.$$

By Lemma 3 and condition (C7), there exists a constant $c > 0$ such that

$$\|\bar{\mathbf{l}}(\widetilde{\boldsymbol{\gamma}}) - \bar{\mathbf{l}}(\widetilde{\boldsymbol{\gamma}}^{(o)})\|_2 = \|(\mathbf{V}_n^*)^\top \mathbf{V}_n^* (\widetilde{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}^{(o)})\|_2 = \|(\mathbf{V}_{n,A}^*)^\top \mathbf{V}_{n,A}^* (\widetilde{\boldsymbol{\gamma}}_A - \widetilde{\boldsymbol{\gamma}}_A^{(o)})\|_2 = \|\mathbf{W}_n\|_2 \le c a_n \lambda_n^2 N_n = o(n r_n \lambda_n).$$

Then Lemma 5 and condition (C7) entail that

$$\Pr\left\{ \max_{d,k,h \in \mathbb{J}_{dk}^{\mathbb{C}}} \|\bar{\boldsymbol{\ell}}_{dk,h}(\widetilde{\boldsymbol{\gamma}})\|_2 \ge n r_n \lambda_n \right\} \to 0.$$

Therefore, $\widetilde{\gamma}$ satisfies the KKT condition with probability approaching to 1. Consequently, the solution to (3) with the SCAD penalty is asymptotically equivalent to the oracle estimator. Then Theorem 2 follows from the triangle inequality and condition (C7). □

References

[1] I. Cribben, R. Haraldsdottir, L.Y. Atlas, T.D. Wager, M.A. Lindquist, Dynamic connectivity regression: determining state related changes in brain connectivity, NeuroImage 61 (2012) 907–920.
[2] C. de Boor, A Practical Guide to Splines, New York: Springer, 2001.
[3] J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[4] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[5] J. Grinband, T.D. Wager, M.A. Lindquist, V.P. Ferrera, J. Hirsch, Detection of time-varying signals in event-related fMRI designs, NeruroImage 43 (2008) 509–520.
[6] Y.T. Guan, On consistent nonparametric intensity estimation for inhomogeneous spatial point processes, J. Amer. Statist. Assoc. 103 (2008) 1238–1247.
[7] Y.T. Guan, Y. Shen, A weighted estimating equation approach for inhomogeneous spatial point processes, Biometrika 97 (2010) 867–880.
[8] T. Hastie, R. Tibshirani, Varying-coefficient models, J. R. Stat. Soc. Ser. B Stat. Methodol. 55 (1993) 757–796.
[9] X. He, W.K. Fung, Z.Y. Zhu, Robust estimation in generalized partial linear models for clustered data, J. Amer. Statist. Assoc. 472 (2005) 1176–1184.
[10] J.H. Hu, P. Wang, A. Qu, Estimating and identifying unspecified correlation structure for longitudinal data, J. Comput. Graph. Statist. 24 (2015) 455–476.
[11] J. Huang, C.O. Wu, L. Zhou, Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements, Biometrika 89 (2002) 111–128.
[12] J.Z. Huang, C.O. Wu, L. Zhou, Polynomial spline estimation and inference for varying coefficient models with longitudinal data, Statist. Sinica 14 (2004) 763–788.
[13] G. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, Ann. Statist. 37 (2009) 2083–2108.
[14] M. Jun, M.L. Stein, Nonstationary covariance models for global data, Ann. Appl. Stat. 2 (2008) 1271–1289.
[15] H. Kang, H. Ombao, C. Linkletter, N. Long, D. Badre, Spatio-spectral mixed effects model for functional magnetic resonance imaging data, J. Amer. Statist. Assoc. 107 (2012) 568–577.
[16] C. Lam, J. Fan, Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation, Ann. Statist. 37 (2009) 4254–4278.
[17] B. Li, M.G. Genton, M. Sherman, A nonparametric assessment of properties of space-time covariance functions, J. Amer. Statist. Assoc. 102 (2007) 736–744.
[18] Y.H. Li, Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation, Biometrika 98 (2011) 355–370.
[19] Y.H. Li, Y. Guan, Functional principal component analysis of spatial-temporal point processes, with applications to disease surveillance, J. Amer. Statist. Assoc. 109 (2014) 1205–1215.
[20] B. Li, H. Zhang, An approach to modeling asymmetric multivariate spatial covariance structures, J. Multivariate Anal. 102 (2011) 1445–1453.
[21] K.Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[22] M.A. Lindquist, The statistical analysis of fMRI data, Statist. Sci. 23 (2008) 439–464.
[23] M.A. Lindquist, Functional causal mediation analysis with an application to brain connectivity, J. Amer. Statist. Assoc. 107 (2012) 1297–1309.
[24] H. Miller, P. Hall, Local Polynomial Regression and Variable Selection, IMS Collections 6, Borrowing Strength: Theory Powering Applications –a Festschrift for Lawrence D. Brown 1 2010 pp. 216–233.
[25] A. Qu, R.Z. Li, Quadratic inference functions for varying coefficient models with longitudinal data, Biometrics 62 (2006) 379–391.
[26] A. Qu, B. Lindsay, B. Li, Improving generalised estimating equations using quadratic inference functions, Biometrika 87 (2000) 823–836.
[27] H. Sang, M. Jun, J.Z. Huang, Covariance approximation for large multivariate spatial datasets with an application to multiple climate model errors, Ann. Appl. Stat. 5 (2011) 2519–2548.
[28] L.L. Schumaker, Spline Function, New York: Wiley, 1981.
[29] M.L. Stein, Space-time covariance functions, J. Amer. Statist. Assoc. 100 (2005) 310–321.
[30] M.L. Stein, Z.Y. Chi, L.J. Welty, Approximating likelihoods for large spatial data sets, J. R. Stat. Soc. Ser. B Stat. Methodol. 66 (2004) 275–296.
[31] R.J. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996) 267–288.
[32] R.J. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 91–108.
[33] P. Wang, J. Zhou, A. Qu, Correlation structure selection for longitudinal data with diverging cluster size, Canad. J. Statist. 44 (2016) 343–360.
[34] R.W.M. Wedderburn, Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method, Biometrika 61 (1974) 439–447.

[35] F. Wei, J. Huang, H. Li, Variable selection and estimation in high-dimensional varying coefficient models, Statist. Sinica 21 (2011) 1515–1540.
[36] L. Xue, Variable selection in additive models, Statist. Sinica 19 (2009) 1281–1296.
[37] L. Xue, A. Qu, Variable selection in high-dimensional varying-coefficient models with global optimality, J. Mach. Learn. Res. 13 (2012) 1973–1998.
[38] L. Xue, L. Yang, Additive coefficient modeling via polynomial spline, Statist. Sinica 16 (2006) 1423–1446.
[39] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. (2010) 894–942.
[40] J.H. Zhou, A. Qu, Informative estimation and selection of correlation structure for longitudinal data, J. Amer. Statist. Assoc. 107 (2012) 701–710.
[41] S. Zhou, X. Shen, D.A. Wolfe, Local Asymptotiic for Regression Splines and Confidence Regions, Ann. Statist. 26 (1998) 1760–1782.
[42] J.H. Zhou, N.Y. Wang, N. Wang, Functional linear model with zero-value coefficient function at sub-regions, Statist. Sinica 23 (2013) 25–50.
[43] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.