Statistica Sinica Preprint No: SS-2017-0218	
Title	TIME-VARYING ESTIMATION AND DYNAMIC MODEL SELECTION WITH AN APPLICATION OF NETWORK DATA
Manuscript ID	SS-2017-0218
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0218
Complete List of Authors	Lan Xue Xinxin Shu and Annie Qu
Corresponding Author	Annie Qu
E-mail	anniequ@illinois.edu
Notice: Accepted version subject to English editing.	

TIME-VARYING ESTIMATION AND DYNAMIC MODEL SELECTION WITH AN APPLICATION OF NETWORK DATA

Lan Xue, Xinxin Shu and Annie Qu

Oregon State University, Merck and University of Illinois at Urbana-Champaign

Abstract: In many biomedical and social science studies it is important to identify and predict the dynamic changes of associations among network data over time. We propose a varying-coefficient model to incorporate time-varying network data, and impose a piecewise-penalty function to capture local features of the network associations. The advantages of the proposed approach are that it is semi-parametric and therefore flexible in modeling dynamic changes of association for network data problems, and capable of identifying the time regions when dynamic changes of associations occur. To achieve sparsity of network estimation at local time intervals, we implement a group penalization strategy involving overlapping parameters among different groups. However, this imposes great challenges in the optimization process for handling large-dimensional network data observed at many time points. We develop a fast algorithm, based on the smoothing proximal gradient method, which is computationally efficient and accurate. We illustrate the proposed method through simulation studies and children's attention deficit hyperactivity disorder fMRI data, and show that the proposed method and algorithm efficiently recover dynamic network changes over time.

Key words and phrases: B-spline; Dynamic network; Model selection consistency; Proximal gradient method; Varying-coefficient model.

1. Introduction

In social science, genomic, environmental and biomedical studies, it is scientifically important to identify and predict associations and interactions among genes, spatial locations or social structures effectively. Network modeling (e.g., Kolaczyk 2009) can effectively quantify the associations among variables. Our method is motivated by a children's attention deficit hyperactivity disorder study, where the data can be obtained from the ADHD-200 sample initiative website `http://fcon_1000.projects.nitrc.org/indi/adhd200/`. The test samples contain fMRI data from different regions of interest of ADHD children's brains, which are repeatedly measured at many time points. We are interested in identifying associations and interactions among different regions of interest of the brain over time so we can better understand how ADHD patients' brains function.

Figure 1 illustrates the dynamic changes of associations among several regions of interest of a brain over three time-points. We are interested in extracting the underlying signals of associations

through modeling responses of brain activities over time. This can be formulated as a time-varying network problem, where the regions of interest are variables or nodes in the network, and the associations among regions of interest represent edges connecting nodes of the network.

Recent development on network modeling includes high-dimensional graphical models by Meinshausen and Bühlmann (2006); Friedman et al. (2007); and Peng et al. (2009). The central idea of these approaches is to estimate the precision matrix or the inverse of the covariance matrix which provides a conditional correlation interpretation among variables in the graph, where zero partial correlation implies pairwise conditional independence. In addition, Shen et al. (2012) and Zhu et al. (2013) develop simultaneous grouping pursuit and feature selection for high-dimensional graphs. For multiple graphs, Guo et al. (2011) jointly estimate graphical models to capture the dependence among multiple graphs and their common structure, and Zhu et al. (2014) propose the maximum penalized likelihood approach to model structural changes over multiple graphs to incorporate dependency among interacting units.

Most of the existing literature targets the network data problem observed at one-time-point only. However, networks can be observed at multiple time-points where the dynamic changes of associations is of scientific interest and requires quantification. For example, in gene expression data, functional magnetic resonance imaging (fMRI), and social network data, it is common that associations can change over time, and therefore it is important to model and estimate the dynamic changes of the network structure.

Modeling time-varying network data could be statistically and computationally challenging as the network structures over time could be quite complex, involve large-dimensional parameter estimation, and be computationally highly intensive with high-dimensional matrix operations. Existing approaches for time-course network data include linear mixed-effect modeling to incorporate temporal correlation (Shojaie and Michailidis 2010), the kernel-reweighted logistic regression method for time-evolving network structure (Song et al. 2009, Kolar et al. 2010), and time-varying Markov random fields (Kolar and Xing 2009). However, these approaches are mainly for the estimation of time-varying networks, and are not designed for model selection to capture the changes of associations in local time regions.

We propose a dynamic network model to capture the changes of associations through a varying-coefficient model (Hastie and Tibshirani 1993, Huang et al. 2002, Cheng et al. 2016). The modeling for dynamics of partial correlations is semi-parametric and therefore flexible in modeling the nonlinear changes of coefficients. In addition, we propose a one-step penalized polynomial spline method to

detect zero regions in the varying coefficients. Therefore, we are able to locate the time regions when dynamic changes of associations occur. This is applicable to identifying the changes of associations among different regions of interest over time as in the example of fMRI data for ADHD patients, which could be potentially useful for detecting dynamic changes in brain functions.

The one-step penalized polynomial spline method proposed in this paper is very different from the penalization methods (Xue 2009, Wei, Huang and Li 2011, Xue and Qu 2012) recently developed for variable selection in semi-parametric models. Those variable selection approaches are developed to determine if a non-parametric function is zero in the entire region. Therefore a L_2 norm of the spline coefficients is penalized to shrink a function to zero on the entire region. However, the one-step penalized polynomial spline method in this paper aims to detect local zero regions in the varying coefficients, therefore to locate the time regions when dynamic changes of associations occur. We utilizes the local property of the polynomial splines that the spline functions on a given local interval only depend on the neighboring B-spline bases. Therefore, we propose to penalize only those coefficients relevant to a given local interval in a group-wise fashion. This new form of penalization raises challenges in both computation and theory development, that we will discuss in details in Sections 3 and 4.

In order to achieve sparsity for network data at local time intervals, we propose a piecewise penalized loss function incorporating the local features of the varying-coefficient models in dynamic modeling. The piecewise penalization strategy involves overlapping spline-coefficient parameters among different penalty groups. However, the popular coordinate-wise descent algorithm cannot be applied in our optimization. We propose an alternative algorithm which is computationally efficient and accurate based on the proximal gradient method. The advantage of this approach is that it does not involve large-dimensional matrix inversion, and is capable of handling large-dimensional network data.

One computational challenge we face for time-varying network data is that the volume of this type of data is extremely large, as it includes observations for many nodes over many time points. For example, when the network size is about 100 and observed over 50 time points, the dimension of the matrix operation could reach 10^5 in iteration process. Existing methods for handling time-varying networks mainly target relatively small network sizes with limited time points. Therefore there is a great demand to develop computationally efficient and fast algorithms to solve the large-dimensional time-varying network problem. The proposed group penalization strategy effectively ensures sparsity at local time intervals; however, it brings additional computational cost in the optimization process,

as it requires a high degree of memory storage and matrix operations for solving the dynamic network problem. In theory, it is also more challenging to establish local-feature model selection consistency than global-feature model selection consistency. We show that the proposed method identifies zero estimators in the non-signal time regions, and estimates the partial correlation functions uniformly consistently in the signal regions.

Recent work on dynamic modeling for network change includes Lebre et al.'s (2010) reversible jump MCMC, Zhou et al.'s (2010) time-series model for covariance matrix, Kolar et al. (2009), Kolar and Xing (2011) and Kolar and Xing's (2012) piecewise constant varying-coefficient varying-structure (VCVS) models, and Chen and Leng's (2016) nonparametric model for the dynamic covariance matrix. Our approach is very different from these approaches as we use the penalized polynomial spline function for modeling network change which is able to accommodate many time points with scalable computing cost. This is in contrast to the reversible jump MCMC approach, which is mainly applicable for a limited number of time points, and is also very different from the piecewise constant VCVS approach which models abrupt change instead of smooth change for network structure. Zhou et al.'s (2010) method is based on the penalized maximum likelihood approach where the covariance matrix is estimated through a kernel smoother. However, they have not established the sparsistency property by which all zero parameters are estimated as zero with probability approaching one. In contrast, we establish the sparsistency property for the proposed method, which is quite important for detecting dynamic network structure change. Chen and Leng's (2016) approach is nonparametric in that no assumption is assumed on the covariance matrix, while our method is semi-parametric in that we model each partial correlation function as a semi-parametric varying-coefficient function.

In addition, the development of dynamic brain network models is also quite active. To study neural connectivity disruptions caused by disease pathology, it is important to develop dynamic brain network models which capture the temporal connectivity of brain networks. Current dynamic brain network models include dynamic causal models (DCM) (Friston et al., 2003) and a nonlinear extension of the DCM (Stephan et al., 2008) which builds on the causal neuronal model with dynamic specified input, state, and output variables corresponding to stimulus functions, the neuronal activities or biophysical variables, and the outcomes measured from the brain regions of interest. In addition, Wang et al. (2015) investigate the important role of the dynamic temporal-topological structure of the ADHD brain network using sliding time-window correlation coefficients. Wee et al. (2016) proposed the fused sparse learning algorithm to jointly estimate temporal networks, while encouraging temporally correlated networks to form similar network structures through the fused Lasso (Tibshirani

et al., 2005). Furthermore, Lee et. al. (2011) recover the sparse brain network derived from partial correlations when the sample size is relatively small while the dimension of parameters is high. Wee et al. (2012) also consider a constrained sparse linear regression model using the LASSO penalty when there is a relatively small number of connections among brain networks. However, the sparse network models do not incorporate dynamic changes of the brain network.

Furthermore, the diffusion wavelet has been proposed to analyze time-varying brain networks. It provides a framework to study properties and structures of a graph in the spectral domain and provides multi-resolution and interpretable basis representations for network data. Chung (1997) gave a comprehensive overview of spectral graph theory. Leonardi and Van De Ville (2011) applied spectral graph wavelet transform (SGWT) to brain functional connectivity data. They decomposed fMRI data using the SGWT and used wavelet coefficients to understand the connectivity of the network. However, this connectivity can only be interpreted in a specific frequency band. Kim et al. (2013) applied the diffusion wavelet to conduct multi-resolution analysis on brain networks and compared connectivity differences between healthy and bipolar patients. All these works focused on representing information contained in a graph via a few interpretable wavelet bases, which capture structure differences in brain networks. In general, the use of the diffusion wavelet is to reduce dimensionality while appropriately incorporating the topology information in the network. In contrast, our work aims to model pairwise connectivity of the network. For future research, one may first use our method to estimate network connectivities, followed by a multi-resolution analysis using the diffusion wavelet to understand the differences in such networks.

The paper is organized as follows. Section 2 proposes the penalized polynomial spline method for time-varying network data. Section 3 provides the smoothing proximal gradient algorithm to capture dynamic changes of network data over time. Section 4 presents asymptotic theory of model selection local consistency. In Section 5, we compare the numerical performance of the proposed smoothing proximal gradient algorithm with other existing approaches. Section 6 illustrates the proposed method for the fMRI data of ADHD patients. The final section provides concluding remarks and a brief discussion.

2. Time-varying networks

In this paper, we focus on time-varying network data and are interested in modeling dynamic changes in its partial correlations or structural changes of the network over time. Both the correlation function and the partial correlation function can be used to characterize associations among variables of interest. We focus on the partial correlation function mainly due to the fact that we are

interested in **conditional** dependence/independence among variables in a network. It measures the direct relationship between two variables while removing the influence of other variables.

Let $\mathbf{y}(t) = \big(y_1(t), \cdots, y_p(t)\big)'$ be a set of time-varying variables observed at time t, and $\{\mathbf{y}(t), t \in \mathbf{I}\}$ be the corresponding continuous stochastic process defined on a compact interval $\mathbf{I}$. Without loss of generality, let $\mathbf{I} = [0, 1]$. Suppose the data consists of n subjects with measurements taken at m discrete time-points $0 \leq t_{k1} < \cdots < t_{km} \leq 1$ for each subject $k = 1, \ldots, n$. For each subject, the observation $\mathbf{y}^k(\mathbf{t}_k) = \big(\mathbf{y}^k(t_{k1}), \cdots, \mathbf{y}^k(t_{km})\big)'$ is a discrete realization of the continuous stochastic process $\{\mathbf{y}(t), t \in \mathbf{I}\}$ at m subject-specific time points $\mathbf{t}_k = (t_{k1}, \ldots, t_{km})$. Here $\mathbf{y}^k(t_{ku}) = \big(y_1^k(t_{ku}), \cdots, y_p^k(t_{ku})\big)'$ for $u = 1, \ldots, m$ are p variables observed at time t_{ku} for the kth subject.

Let $\boldsymbol{\rho}(t) = \{\rho^{12}(t), \cdots, \rho^{(p-1)p}(t)\}'$ be the partial correlation function of $\mathbf{y}(t)$. Suppose each partial coefficient function $\rho^{ij}(t)$ varies in time smoothly. We can apply the polynomial spline to approximate the time-varying coefficients since it provides a good approximation of any smooth function, even with a small number of knots. Let $\{\nu_h\}_{h=1}^{N_n}$ be N_n interior knots within the interval $[0, 1]$ and Υ be a partition of the interval $[0, 1]$ with N_n knots. That is $\Upsilon_n = \{0 = \nu_0 < \nu_1 < \cdots < \nu_{N_n} < \nu_{N_n+1} = 1\}$. The polynomial splines of order $q + 1$ are functions with q-degree of polynomials on intervals $[\nu_{h-1}, \nu_h), h = 1, \ldots, N_n$ and $[\nu_{N_n}, \nu_{N_n+1}]$, and $q - 1$ continuous derivatives globally. We denote the space of such spline functions by G_n. Let $\{B_h(\cdot)\}_{h=1}^{J_n}$ be a set of B-spline bases of G_n, where $J_n = N_n + q + 1$ and the function $\rho^{ij}(t)$ for any $1 \leq i < j \leq p$ can be approximated by

$$\rho^{ij}(t) \approx g^{ij}(t) = \sum_{h=1}^{J_n} \beta_h^{ij} B_h(t) = (\beta^{ij})' \mathbf{B}(t),$$

where $\beta^{ij} = \big(\beta_1^{ij}, \ldots, \beta_{J_n}^{ij}\big)'$ is a set of coefficients, and $\mathbf{B}(t) = (B_1(t), \ldots, B_{J_n}(t))'$ are B-spline bases. In practice, different B-spline bases can be used to approximate different $\rho^{ij}(t)$. For simplicity, the same set of B-spline bases is used for different partial correlation functions in this paper.

In addition to polynomial splines, other basis functions can also be used to approximate unknown functions including wavelet and trigonometric polynomials. Sections 2.5 and 2.6 of Fan and Gijbels (1996) provide a review of the basis choices. The reason we choose the polynomial spline is due to its sound numerical properties as well as excellent approximation powers. Given a sufficient number of knots, any continuous function can be approximated arbitrarily well by polynomial splines under the assumption that it is reasonably smooth. However, in general, polynomial splines cannot approximate functions with discontinuities and rapid variations sufficiently well. For discontinuous or rapidly

varying functions, other basis functions such as the wavelet might be more suitable.

Suppose $\mathbf{y}(t)$ has mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}(t)$. Denote the concentration matrix $\mathbf{\Sigma}^{-1}(t)$ by $(\sigma^{ij}(t))_{p\times p}$. Then one can express $y_i(t)$ by a varying coefficient model as

$$y_i(t) = \sum_{j\neq i} \beta_{ij}(t) y_j(t) + \varepsilon_i(t), \tag{2.1}$$

with $\beta_{ij}(t) = \rho^{ij}(t)\sqrt{\sigma^{jj}(t)/\sigma^{ii}(t)}$, and $Var(\varepsilon_i(t)) = 1/\sigma^{ii}(t)$. The errors $\varepsilon_i(t)$ can be correlated over time. However, in the following, the longitudinal correlation is not incorporated, and instead we assume that $\varepsilon_i(t)$ is independent over time. We focus on the development of methods to identify the local sparsity of the coefficient functions $\{\beta_{ij}(t)\}$ over time. In the traditional polynomial spline estimation, one replaces $\rho^{ij}(t)$ with $g^{ij}(t)$, and estimates the spline coefficients $\beta = \{\beta^{ij}, 1 \leq i < j \leq p\}$ by minimizing the weighted sum of squares in (2.2). The advantages of spline approximation for the time-varying coefficient model are that it is computationally fast and efficient.

In this paper we are interested in locally sparse estimators of the partial correlations that can characterize dynamic changes of network associations over time. The B-spline basis function has a desirable local property. For any interval constructed by two consecutive knots, denote as (ν_{h-1}, ν_h) for $1 \leq h \leq N_n + 1$. If $t \in (\nu_{h-1}, \nu_h)$, the spline function $g^{ij}(t)$ is only affected by basis functions $B_h, \ldots, B_{h+q}$. Therefore, the spline function $g^{ij}(t)$ is locally zero within the interval (ν_{h-1}, ν_h), if and only if the spline coefficients $\boldsymbol{\gamma}_h^{ij} = (\beta_h^{ij}, \ldots, \beta_{(h+q)}^{ij})'$ are all zero. In addition, the whole region $[0, 1]$ can be divided into $N_n + 1$ intervals by the spline knots. Therefore, we penalize the group of spline coefficients associated with each local interval $[\nu_{h-1}, \nu_h]$ in a group-wise fashion. Consequently, this provides locally sparse spline estimators $\widetilde{\rho}_{ij}(t)$ which can be completely zero on certain time intervals spanned by the knot sequence.

We propose the following piecewise penalized loss function to achieve sparsity for the network data:

$$\begin{aligned} PL(\boldsymbol{\beta}, \boldsymbol{\sigma}, \mathbf{t}, \mathbf{y}) \;=\; & \frac{1}{2nm} \sum_{k=1}^{n} \sum_{i=1}^{p} \sum_{u=1}^{m} w_{i_{ku}} \left(y_i^k(t_{ku}) - \sum_{j\neq i}^{p} \sum_{h=1}^{J_n} \beta_h^{ij} B_h(t_{ku}) \sqrt{\frac{\sigma^{jj}(t_{ku})}{\sigma^{ii}(t_{ku})}} y_j^k(t_{ku}) \right)^2 \\ & + \sum_{i<j}^{p} \sum_{h=1}^{N_n+1} P_{\lambda_n}(\|\boldsymbol{\gamma}_h^{ij}\|), \end{aligned} \tag{2.2}$$

where $\mathbf{y} = \{\mathbf{y}^k(\mathbf{t}_k)\}_{k=1}^n$, $\boldsymbol{\beta} = (\beta_1^{1,2}, \ldots, \beta_{J_n}^{1,2}, \ldots, \beta_1^{p-1,p}, \ldots, \beta_{J_n}^{p-1,p})'$ is a $p(p-1)J_n/2$-dimensional

spline coefficient, $\boldsymbol{\sigma} = \{\sigma^{ii}(\mathbf{t})\}_{i=1}^{p}$ with $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_n)'$, and $w_{i_{ku}}$ are nonnegative weights and typically can be chosen as $\sigma^{ii}(t_{ku})$. In addition, $\|\cdot\|$ is the vector L_2-norm. Note that in contrast to Peng et al.'s (2009) loss function, both the weights and the components in the concentration matrix are varying over time.

The first term of (2.2) is the weighted sum of squares, and the second term P_{λ_n} of (2.2) is the penalty function which can be chosen from LASSO, SCAD or the adaptive LASSO described in subsection 3.1. The performance of the penalty function crucially depends on the tuning parameter λ_n, whose selection will be discussed in subsection 3.2. Intuitively, if $\|\gamma_h^{ij}\|$ is shrunk towards zero, then all elements of γ_h^{ij} are zero and the spline function $g^{ij}(t)$ is locally zero on the corresponding interval. The penalty term in (2.2) is different from the typical penalty for global model selection in semi-parametric models, such as those proposed in Xue (2009), and Xue and Qu (2012). Here we incorporate the local features of varying-coefficient models and ensure local sparsity of the dynamic modeling. Zhou et al. (2013) incorporated similar idea to detect zero sub-regions for functional coefficients in functional linear regression model through a two-step procedure.

Both $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ are unknown parameters but $\boldsymbol{\beta}$ is the main parameter of our interest. To estimate $\boldsymbol{\beta}$ in the penalized loss (2.2), $\boldsymbol{\sigma}$ needs to be specified and a two-step iterative procedure will be proposed in the algorithm in the next section.

Let $\mathbf{y}_{iu} = (y_i^1(t_{1u}), \ldots, y_i^n(t_{nu}))'$, $\widetilde{\mathbf{y}}_{iu} = \sqrt{\frac{w_{iu}}{nm}}\mathbf{y}_{iu}$, $\widetilde{\mathbf{y}}_i = (\widetilde{\mathbf{y}}_{i1}', \ldots, \widetilde{\mathbf{y}}_{im}')'$, and $\mathcal{Y}_n = (\widetilde{\mathbf{y}}_1', \ldots, \widetilde{\mathbf{y}}_p')'$ be a nmp-dimensional vector. Let $\mathcal{X}_n = (\widetilde{\mathbf{x}}_{(1,2)}', \ldots, \widetilde{\mathbf{x}}_{(p-1,p)}')$ be a $(nmp) \times \{p(p-1)J_n/2\}$-dimensional matrix, with $\widetilde{\mathbf{x}}_{(i,j)} = (\mathbf{0}_1, \ldots, \mathbf{0}_{i-1}, \mathbf{z}_{(i,j)}^j, \mathbf{0}_{i+1}, \ldots, \mathbf{0}_{j-1}, \mathbf{z}_{(i,j)}^i, \ldots, \mathbf{0}_p)'$, where $\mathbf{0}_k = \{0\}_{J_n \times nm}$, and $\mathbf{z}_{(i,j)}^j = \left(\mathbf{z}_{(i,j),1}^j, \ldots, \mathbf{z}_{(i,j),m}^j\right)'$, with

$$\mathbf{z}_{(i,j),u}^j = \left(\mathbf{B}(t_{1u})\sqrt{\frac{\widetilde{\sigma}^{jj}(t_{1u})}{\widetilde{\sigma}^{ii}(t_{1u})}}y_j^1(t_{1u}), \ldots, \mathbf{B}(t_{nu})\sqrt{\frac{\widetilde{\sigma}^{jj}(t_{nu})}{\widetilde{\sigma}^{ii}(t_{nu})}}y_j^n(t_{nu})\right),$$

for $u = 1, \ldots, m$, and $\widetilde{\sigma}^{ii}(t_u) = \sigma^{ii}(t_u)/w_{iu}$. Then the corresponding loss function (2.2) is equivalent to

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}, \mathcal{Y}_n) = \frac{1}{2}\|\mathcal{Y}_n - \mathcal{X}_n\boldsymbol{\beta}\|^2 + \sum_{i<j}^{p}\sum_{h=1}^{N_n+1} P_{\lambda_n}(\|\boldsymbol{\gamma}_h^{ij}\|). \tag{2.3}$$

Let $\hat{\boldsymbol{\beta}}$ be the minimizer of object function (2.2) or (2.3). Then the resulting estimator for the partial correlation function $\rho^{ij}(t)$ is defined as $\hat{\rho}^{ij}(t) = \left(\hat{\beta}^{ij}\right)^T \mathbf{B}(t)$.

3. Implementations

3.1 Algorithms

In this section, we propose an algorithm to obtain an optimal solution for the objective function (2.3). Let the penalty function $P_{\lambda_n}(\|\gamma_h^{ij}\|)$ in (2.3) follow the adaptive Lasso penalty (Tibshirani 1996, Zou 2006), that is, $P_\lambda(\|\gamma_h^{ij}\|) = \lambda_n \tau_h^{ij} \|\gamma_h^{ij}\|$, where $\tau_h^{ij} = 1/\|\tilde{\gamma}_h^{ij}\|^r$ with $r > 0$ and $\tilde{\gamma}_h^{ij}$ is a consistent estimator of γ_h^{ij}. So the penalty term can be considered as an adaptive group LASSO with overlapping groups. When the groups overlap, if one group is shrunk to zero, all the coefficients in this group shrink to zero, even though some coefficients in this group also belong to other nonzero-coefficient groups. The solution space and theoretical properties of the group LASSO with overlaps are discussed in Jenatton et al. (2011) and Obozinski et al. (2011), which indicate that traditional algorithms for LASSO cannot be directly applied to the penalized loss function in (2.2).

However, since the dual norm of the L_2-norm is still the L_2-norm, the L_2-norm $\boldsymbol{\gamma}_h^{ij}$ can be formulated as $\max_{\|\boldsymbol{\alpha}_h^{ij}\| \le 1} (\boldsymbol{\alpha}_h^{ij})' \boldsymbol{\gamma}_h^{ij}$, where $\boldsymbol{\alpha}_h^{ij} \in R^{(q+1)}$ is an auxiliary vector associated with $\boldsymbol{\gamma}_h^{ij}$. A similar transformation and its properties have been discussed in Chen et al. (2012), Jacob et al. (2009) and Obozinski et al. (2011). Let $Q = \{\boldsymbol{\alpha} | \|\boldsymbol{\alpha}_h^{ij}\| \le 1, 1 \le i < j \le p, h = 1, \cdots, N_n + 1\}$. We can rewrite the group adaptive LASSO penalty for the overlapping parameters in (2.2) as follows:

$$g_0(\boldsymbol{\beta}) = \lambda_n \sum_{i<j}^{p} \sum_{h=1}^{N_n+1} \tau_h^{ij} \|\boldsymbol{\gamma}_h^{ij}\| = \max_{\boldsymbol{\alpha} \in \mathbf{Q}} \sum_{i<j}^{p} \sum_{h=1}^{N_n+1} \lambda_n \tau_h^{ij} (\boldsymbol{\alpha}_h^{ij})' \boldsymbol{\gamma}_h^{ij} = \max_{\boldsymbol{\alpha} \in \mathbf{Q}} \boldsymbol{\alpha}' C \boldsymbol{\beta}, \tag{3.1}$$

where $C \in R^{[(q+1)(N_n+1)p(p-1)/2] \times [p(p-1)J_n/2]}$ is an indicator matrix with the element defined as

$$C_{(k,l)} = \begin{cases} \lambda_n \tau_h^{ij} & k = (r-1)(N_n+1)(q+1) + (h-1)(q+1) + v, l = (r-1)J_n + (h-1) + v \\ 0 & \text{otherwise} \end{cases},$$

where $r = (i-1)(p-i+2) + (j-i-1)$ and $v = 1, \cdots, (q+1)$. Note that C is a very sparse matrix with only one non-zero element in each row, and therefore only requires a relatively small amount of memory storage in the optimization procedure. Through the transformation, the group penalization terms no longer present overlapping parameters.

However, this introduces a new problem, in that the penalty function $g_0(\boldsymbol{\beta})$ in (3.1) is a non-smooth function of $\boldsymbol{\beta}$. To circumvent this problem, we need to build a smooth function to approximate

$g_0(\boldsymbol{\beta})$. Let $D = \max_{\boldsymbol{\alpha}\in\mathbf{Q}} \|\boldsymbol{\alpha}\|^2/2$ and

$$g_\mu(\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}\in\mathbf{Q}} \left(\boldsymbol{\alpha}' C\boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|^2\right), \tag{3.2}$$

where μ is the tolerance parameter. Then $g_\mu(\boldsymbol{\beta})$ is a quadratic approximation for $g_0(\boldsymbol{\beta})$ with the maximum difference of μD. That is,

$$g_0(\boldsymbol{\beta}) - \mu D \le g_\mu(\boldsymbol{\beta}) \le g_0(\boldsymbol{\beta}).$$

In order to control the maximum difference, we choose the tolerance level $\epsilon = \mu D$, or equivalently $\mu = \epsilon/D$. Consequently, the loss function in (2.3) can be approximated by

$$\widetilde{PL}(\mu, \boldsymbol{\beta}, \boldsymbol{\sigma}) = \frac{1}{2}\|\mathcal{Y}_n - \mathcal{X}_n\boldsymbol{\beta}\|^2 + g_\mu(\boldsymbol{\beta}).$$

To minimize the loss function $\widetilde{PL}(\mu, \boldsymbol{\beta})$, we need to calculate the gradient of $\widetilde{PL}(\mu, \boldsymbol{\beta})$. For any $\mu > 0$, $g_\mu(\boldsymbol{\beta})$ is convex and continuously differentiable and the corresponding gradient function $\nabla g_\mu(\boldsymbol{\beta})$ is $C'\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$ is the optimal solution in (3.2). Let $\boldsymbol{u}_h^{ij} = \lambda_n \tau_h^{ij} \boldsymbol{\gamma}_h^{ij}/\mu$ and the closed form of $\boldsymbol{\alpha}^*$ can be expressed as

$$(\boldsymbol{\alpha}_h^{ij})^* = \begin{cases} \dfrac{\boldsymbol{u}_h^{ij}}{\|\boldsymbol{u}_h^{ij}\|}, & \text{if } \left\|\boldsymbol{u}_h^{ij}\right\| > 1 \\ \boldsymbol{u}_h^{ij}, & \text{if } \left\|\boldsymbol{u}_h^{ij}\right\| \le 1 \end{cases}. \tag{3.3}$$

Therefore the partial derivative $\nabla\widetilde{PL}(\mu, \boldsymbol{\beta}, \boldsymbol{\sigma})$ with respect to $\boldsymbol{\beta}$ can be calculated as $\mathcal{X}_n' (\mathcal{X}_n\beta - \mathcal{Y}_n) + C'\alpha^*$. Moreover, $\nabla\widetilde{PL}(\mu, \boldsymbol{\beta}, \boldsymbol{\sigma})$ is Lipschitz-continuous with the Lipschitz constant

$$M = \lambda_{\max}(\mathcal{X}_n'\mathcal{X}_n) + \frac{\|C\|^2}{\mu},$$

where $\lambda_{\max}$ is the largest eigenvalue of $(\mathcal{X}_n)'\mathcal{X}_n$ and $\|C\| = \max_{\|\boldsymbol{\alpha}\|\le 1} \|C\boldsymbol{\alpha}\|$. The proximal operator can be defined as

$$Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}', \boldsymbol{\sigma}) = \left\{\widetilde{PL}(\mu, \boldsymbol{\beta}', \boldsymbol{\sigma}) + \nabla PL(\mu, \boldsymbol{\beta}', \boldsymbol{\sigma})(\boldsymbol{\beta} - \boldsymbol{\beta}') + \frac{M}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|^2\right\},$$

and $\boldsymbol{\beta}$ can be updated at the $(l+1)$th iteration by applying the proximal gradient algorithm through

$$\begin{aligned}\boldsymbol{\beta}^{(l+1)} &= arg\min_{\boldsymbol{\beta}} Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}) \\ &= arg\min_{\boldsymbol{\beta}} \left\{ \widetilde{PL}(\mu, \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}^{(l)}) + \nabla PL(\mu, \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}^{(l)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(l)}) + \frac{M}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(l)}\|^2 \right\} \quad (3.4)\end{aligned}$$

Convergence is guaranteed since the inequality $\widetilde{PL}(\mu, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\sigma}^{(l)}) \leq Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}^{(l)})$ holds for each iteration. It is not difficult to check if the inequality holds, and details are discussed in Chen et al. (2012). The above penalization strategy is able to achieve sparsity corresponding to the group parameters $\boldsymbol{\gamma}_h$; however, it does not guarantee the sparsity of each element in $\hat{\boldsymbol{\beta}}$ obtained from (3.4). Alternatively, we can set $\beta_h^{ij} = 0$ if the $\|\beta_h^{ij}\| < \epsilon^*$ for a small tolerance level ϵ^*. For σ, if each subject is observed at the same time over m time points, i.e. $t_{ku} = t_u$ for any $k = 1, \ldots, n$ and $u = 1, \ldots, m$, then each component of $\boldsymbol{\sigma}^{(l+1)} = \left\{ \left((\sigma^{11})^{(l+1)}(t_u), \cdots, (\sigma^{pp})^{(l+1)}(t_u) \right) \right\}_{u=1}^{m}$ at the $l+1$-th iteration can be updated by

$$\frac{1}{(\sigma^{ii})^{(l+1)}(t_u)} = \frac{1}{n}\sum_{k=1}^{n}\left(y_i^k(t_u) - \sum_{j\neq i}^{p}\sum_{h=1}^{J_n}(\beta_h^{ij})^{(l)} B_h^{ij}(t_u)\sqrt{\frac{(\sigma^{jj})^{(l)}(t_u)}{(\sigma^{ii})^{(l)}(t_u)}} y_j^k(t_u) \right)^2, \quad (3.5)$$

and the weight component for the ith subject is $w_{iu}^{(l+1)} = (\sigma^{ii})^{(l+1)}$. If each subject is observed at the different m time points, one can update $(\sigma^{ii})^{(l+1)}(t)$ using a polynomial spline estimation method. Let $\hat{\varepsilon}_i^2(t_{ku}) = \left(y_i^k(t_{ku}) - \sum_{j\neq i}^{p}\sum_{h=1}^{J_m}(\beta_h^{ij})^{(l)} B_h^{ij}(t_{ku})\sqrt{\frac{(\sigma^{jj})^{(l)}(t_{ku})}{(\sigma^{ii})^{(l)}(t_{ku})}} y_j^k(t_{ku}) \right)^2$. For each $i = 1, \ldots, p$, one can estimate $\sigma^{ii}(t)$ by a polynomial spline regression using $\{1/\hat{\varepsilon}_i^2(t_{ku})\}_{k=1,u=1}^{n,m}$ as the response variables, and the spline basis generated on time points $\{(t_{ku})\}_{k=1,u=1}^{n,m}$ as explanatory variables. We summarize the algorithm as follows.

Algorithm 1 Proximal gradient algorithm for estimating partial correlation networks

Input: Set desired tolerance levels ϵ and ϵ^* (set to be 10^{-3}), obtain $\mu = \epsilon/D$ and matrix C, and calculate the step size M; initialize the parameters $\boldsymbol{\beta}, \boldsymbol{\sigma}$ as $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\sigma}^{(0)}$.

Output: $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}$.

1: Compute $\boldsymbol{\alpha}^*$ according to (3.3) and calculate $\nabla\widetilde{PL}(\boldsymbol{\beta}^{(l)}, \mu) = \mathcal{X}_n'(\mathcal{X}_n\boldsymbol{\beta}^{(l)} - \mathcal{Y}_n) + C'\boldsymbol{\alpha}^*$;

2: Obtain $\boldsymbol{\beta}^{(l+1)}$ by minimizing (3.4), i.e., $\boldsymbol{\beta}^{(l+1)} = arg\min_{\boldsymbol{\beta}} Q_L(\boldsymbol{\beta}^{(l)}, \boldsymbol{\beta})$, and set the elements in $\boldsymbol{\beta}^{(l+1)}$ less than ϵ^* as zero;

3: Update $\boldsymbol{\sigma}^{(l+1)}$ and $\mathbf{w}^{(l+1)}$ by calculating (3.5);

4: Return to Step 1 if $\left\| Q_L(\boldsymbol{\beta}^{(l+1)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}^{(l+1)}) - Q_L(\boldsymbol{\beta}^{(l)}, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\sigma}^{(l)}) \right\| > \epsilon$.

We can also apply the alternating direction method of multipliers (ADMM) (Boyd et al. 2011)

to approximate $g_0(\boldsymbol{\beta})$ by $g_\mu(\boldsymbol{\beta})$ in (3.2) as follows. The adaptive LASSO with overlapping group penalty can be solved by a constrained optimization:

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\beta}^*} \frac{1}{2}\|\mathcal{Y}-\mathcal{X}\boldsymbol{\beta}\|^2 + g_\mu(\boldsymbol{\beta}^*), \\ &s.t. \quad \boldsymbol{\beta}=\boldsymbol{\beta}^*. \end{aligned} \tag{3.6}$$

This can be further formulated as a scaled augmented Lagrangian problem:

$$L_\rho = \frac{1}{2}\|\mathcal{Y}-\mathcal{X}\boldsymbol{\beta}\|^2 + g_\mu(\boldsymbol{\beta}^*) + \frac{\kappa}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*+\boldsymbol{\eta}\|_2^2, \tag{3.7}$$

where $\boldsymbol{\eta}$ are dual variables and κ is a scalar and can be preset. Therefore, the ADMM algorithm solving (3.7) leads to three iteration steps for $\boldsymbol{\beta}, \boldsymbol{\beta}^*, \boldsymbol{\eta}$. That is, at the $(l+1)$-th iteration,

$$\begin{aligned} \boldsymbol{\beta}^{(l+1)} &= arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathcal{Y}-\mathcal{X}\boldsymbol{\beta}\|^2 + \frac{\kappa}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^{*(l)}+\boldsymbol{\eta}^{(l)}\|_2^2, \\ \boldsymbol{\beta}^{*(l+1)} &= arg\min_{\boldsymbol{\beta}} g_\mu(\boldsymbol{\beta}^*) + \frac{\kappa}{2}\|\boldsymbol{\beta}^{(l+1)}-\boldsymbol{\beta}^*+\boldsymbol{\eta}^{(l)}\|_2^2, \\ \boldsymbol{\eta}^{(l+1)} &= \boldsymbol{\eta}^{(l)} + \left(\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{*(l+1)}\right). \end{aligned} \tag{3.8}$$

The first minimization problem in (3.8) is easy to solve since the objective function is quadratic. The function $g_\mu(\boldsymbol{\beta}^*)$ in the second minimization is a smoothing function and thus can be approximated by the Taylor expansion at $\boldsymbol{\beta}^{*(l)}$, i.e. $g_\mu(\boldsymbol{\beta}^*) \approx g_\mu(\boldsymbol{\beta}^{*(l)}) + 1/2\nabla g_\mu(\boldsymbol{\beta}^{*(l)})(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*(l)})$. Thus $\nabla g_\mu(\boldsymbol{\beta}^*) \approx \nabla g_\mu(\boldsymbol{\beta}^{*(l)})/2 = C'\boldsymbol{\alpha}^{*(l)}/2$, where $\boldsymbol{\alpha}^{*(l)}$ can be calculated by (3.3) corresponding to $\boldsymbol{\beta}^{*(l)}$. So the solution $\boldsymbol{\beta}^{*(l+1)} = \boldsymbol{\beta}^{(l+1)} + \boldsymbol{\eta}^{(l)} - \lambda C'\boldsymbol{\alpha}^{*(l)}/(2\kappa)$. The algorithm is summarized as in Algorithm 2:

Algorithm 2 Alternating direction method of multipliers for estimating partial correlation networks

Input: Set desired tolerance levels ϵ, ϵ^* and scalar κ, obtain $\mu = \epsilon/D$ and matrix C; initialize the parameters $\boldsymbol{\beta}, \boldsymbol{\sigma}$ as $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\sigma}^{(0)}$.

Output: $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}$.

1: Compute $\boldsymbol{\alpha}^{*(l)}$ according to (3.3);
2: Obtain $\boldsymbol{\beta}^{(l+1)}, \boldsymbol{\beta}^{*(l+1)}, \boldsymbol{\eta}^{(l+1)}$ according to (3.8), and set the elements in $\boldsymbol{\beta}^{(l+1)}$ less than ϵ^* as zero;
3: Update $\boldsymbol{\sigma}^{(l+1)}$ and $\mathbf{w}^{(l+1)}$ by calculating (3.5);
4: Return to Step 1 if $\left\|\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{*(l+1)}\right\| > \epsilon$.

Both SPG and ADMM provide approximations of (3.1), however, they use different approximation methods and therefore the final solutions from SPG and ADMM are different. The proximal gradient method has the following advantages: (1) we can construct a smoothing approximation to the

objective function, which makes the convergence fast; (2) it does not require large matrix inversion and only involves sparse matrix operations. These could reduce algorithm complexity and improve computational speed significantly. On the other hand, the ADMM requires inverting a matrix, which could lead to infeasible computing when the network size is large. More details are provided in Section 5.

3.2 Tuning parameters selection

The choice of tuning parameters is critical as this determines the performance of the proposed method. Tuning parameter selection for the varying-coefficient model involves two parts. One is the selection of the sequence of knots for the polynomial spline, and the other is the selection of the tuning parameter in the penalty function. For simplicity, we set the number of knots to be the same order of $n^{1/(2q+3)}$, where n is the sample size and q is the order of the polynomial spline. This choice of the number of knots balances between the variance and the squared bias of the polynomial spline estimators (Huang, 1998; Xue and Yang 2006; Huang et al., 2007). One can also use a data-driven knot number which can be selected via a BIC procedure similarly as below. More detailed discussion on knot selection can be found in Huang et al. (2004), Xue et al. (2010) and Xue and Qu (2012). Although, for convenience, we select equally-spaced knots in our numerical studies, our theory is developed under a more general setup which allows for more flexible choices of knot sequence.

In the process of selecting the tuning parameters associated with the penalty function, we use the Bayesian Information Criteria (BIC) procedure, which can be found in the model selection literature (e.g., Qu and Li 2006; Wang, Li and Tsai 2007). Specifically, given the tuning parameters λ_n, denote the estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$, and calculate the estimators $\widehat{\boldsymbol{\sigma}}_{\lambda_n}$ and $\widehat{w}_{\lambda_n}$ through the formula (3.5). Let κ_n be the total number of nonzero elements in $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Then the BIC is given as $BIC(\lambda_n) = nm \log\{MSE(\lambda_n)\} + \kappa_n \log(nm)$, with

$$MSE(\lambda_n) = \frac{1}{nm} \sum_{k=1}^{n} \sum_{i=1}^{p} \sum_{u=1}^{m} \widehat{w}_{iu,\lambda_n} \left(y_i^k(t_{ku}) - \sum_{j \neq i}^{p} \sum_{h=1}^{J_n} \widehat{\beta}_{h,\lambda_n}^{ij} B_h^{ij}(t_{ku}) \sqrt{\frac{\widehat{\sigma}_{\lambda_n}^{jj}(t_{ku})}{\widehat{\sigma}_{\lambda_n}^{ii}(t_{ku})}} y_j^k(t_{ku}) \right)^2 .$$

The optimal turning parameter $\widehat{\lambda}_n$ is selected by minimizing $BIC(\lambda_n)$.

4. Asymptotic theory

In this section we investigate the asymptotic properties of the varying-coefficient estimator $\widehat{\boldsymbol{\rho}}(t)$ based on the polynomial spline approximation. Since one distinct feature of our approach is the

estimation and selection of local features in dynamic network modeling, we will focus on establishing the local-feature model selection consistency of $\widehat{\boldsymbol{\rho}}(t)$. That is, if true $\boldsymbol{\rho}(t)$ is 0 for any given region, the estimator of $\boldsymbol{\rho}(t)$ is 0 with probability approaching 1.

Before presenting the asymptotic properties of the proposed model, we first introduce the following regularity conditions that are required to establish the asymptotic properties.

C1: The weights $\{w_{it}\}_{i=1}^{p}$ are uniformly finite for $t \in I$. That is, there exist positive constants w_0 and w_∞ such that $0 < w_0 \leq \min_i\{w_{it}\} \leq \max_i\{w_{it}\} \leq w_\infty < \infty$ for any $t \in I$.

C2: There exists a constant c such that $\max_{1\leq i\leq p} \sup_{t\in\mathbf{I}} |\widehat{\boldsymbol{\sigma}}^{ii}(t) - \boldsymbol{\sigma}^{ii}(t)| \leq c\sqrt{\frac{log(nm)N_n}{nm}}$ holds with probability approaching to 1 as sample size $n \to \infty$.

C3: We assume that for any $t \in \mathbf{I}$, $\mathbf{y}(t)$ has mean $\mathbf{0}$ and covariance matrix $\Sigma(t)$ whose eigenvalues are assumed to be uniformly bounded for $t \in I$. That is, $0 < \inf_{t\in\mathbf{I}} \lambda_{min}(\Sigma(t)) \leq \sup_{t\in\mathbf{I}} \lambda_{max}(\Sigma(t)) < \infty$ where λ_{min} and λ_{max} are the minimum and maximum eigenvalues of $\Sigma(t)$ respectively. Furthermore, for some sufficiently large $l > 0$, $\sup_{t\in\mathbf{I}} E\left|Y_i\left(t\right)\right|^l < +\infty$, for $i = 1, \ldots, p$.

C4: The observation times $\{t_{ku}\}_{k=1,u=1}^{n,m}$ are independent and follow a distribution $f_T(t)$ on I, and $f_T(t)$ is absolutely continuous and the bounded away from zero and infinity.

C5: For $1 \leq i \neq j \leq p$, the partial correlation function $\rho^{ij}(\cdot)$ has q continuous derivatives with $q \geq 1$.

C6: For $1 \leq i \neq j \leq p$, let $E^{ij} \subset I$ be the null region such that $\rho^{ij}(t) = 0$ if $t \in E^{ij}$ and $\rho^{ij}(t) \neq 0$ if $t \in (E^{ij})^c$. If $E^{ij} \neq \emptyset$, we assume that $E^{ij} = [e_1^{ij}, e_2^{ij}]$ is a closed interval. Let $\dot{\rho}^{ij}(t)$ be the first order derivative of $\rho^{ij}(t)$. We assume there exists a constant $C > 0$ such that $|\dot{\rho}^{ij}(t)| \geq C$ for any $t \in [e_1^{ij} - \epsilon, e_1^{ij}] \cup [e_2^{ij}, e_2^{ij} + \epsilon]$ and a small constant $\epsilon > 0$.

C7: The set of knots denoted as $\Upsilon_n = \{0 = \nu_0 < \nu_1 < \cdots < \nu_{N_n} < \nu_{N_n+1} = 1\}$ is quasi-uniform, i.e., there exists $b > 0$ such that

$$\frac{\max(\nu_{h+1} - \nu_h, h = 0, \cdots, N_n)}{\min(\nu_{h+1} - \nu_h, h = 0, \cdots, N_n)} \leq b.$$

C8: The number of interior knots N_n and tuning parameters λ_n satisfies

$$\lambda_n N_n/\alpha_n \to 0, \lambda_n N_n^2/\alpha_n \to \infty, \lambda_n\sqrt{N_n nm/\log(nm)}/\alpha_n \to \infty,$$

where $\alpha_n = \sqrt{N_n/nm} + N_n^{-1}$.

Condition C1 indicates that the weights are bounded away from 0 and infinity. Condition C2 assumes that there exists a consistent estimator for $\sigma^{ii}(t)$, for each $i = 1, \ldots, p$. Similar conditions of C1 and C2 can also be found in Peng et al. (2009). In the supplementary material, we propose an estimator that meets this condition by kernel smoothing of the residuals of least-square fitting as discussed in the algorithm. Conditions C3, C4, C5 and C7 are standard conditions in the polynomial spline framework, and are required to ensure consistency for spline estimation of the varying coefficient model. Similar conditions can be found in Huang et al. (2002), Xue and Qu (2012), and Wang et al. (2014). Condition C6 is used to separate time regions between zero correlation and nonzero correlation, and thus leads to consistency of the partial correlation estimators.

To present our theoretical results, we first introduce an oracle estimator, which estimates each $\rho^{ij}(t)$ under the assumption that the null regions of each $\rho^{ij}(t)$ are known. It is constructed only for the proof of the asymptotic results, and is not useful for analyzing real data. One notes that, for each end point of the null region $E^{ij} = [e_1^{ij}, e_2^{ij}]$ in condition (C6), there exist knots $\nu_{l_1^{ij}}$ and $\nu_{l_2^{ij}}$ in the knot sequence $\Upsilon = \{0 = \nu_0 < \nu_1 < \cdots < \nu_{N_n} < \nu_{N_n+1} = 1\}$ such that $e_1^{ij} \in [\nu_{l_1^{ij}}, \nu_{l_1^{ij}+1})$ and $e_2^{ij} \in [\nu_{l_2^{ij}-1}, \nu_{l_2^{ij}})$. Let $J_{ij} = \left\{1, \ldots, \nu_{l_1^{ij}} - 2, \nu_{l_2^{ij}} + q + 2, \ldots, J_n\right\}$. An oracle estimator $\widetilde{\beta}^{(o)} = \left\{\widetilde{\beta}_h^{ij(o)}, 1 \le h \le J_n, 1 \le i < j \le p\right\}$ is constructed by taking all coefficients $\widetilde{\beta}_h^{ij(o)} = 0$ for $h = \nu_{l_1^{ij}-1}, \ldots, \nu_{l_2^{ij}} + q + 1$ and estimating the rest of the coefficients by minimizing the sum of the squares

$$\frac{1}{2nm} \sum_{i=1}^{p} \sum_{k=1}^{n} \sum_{u=1}^{m} w_{iu} \left(y_i^k(t_{ku}) - \sum_{j \neq i}^{p} \sum_{h \in J_{ij}} \beta_h^{ij} B_h(t_{ku}) \sqrt{\frac{\widehat{\sigma}^{jj}(t_{ku})}{\widehat{\sigma}^{ii}(t_{ku})}} y_j^k(t_{ku}) \right)^2. \tag{4.1}$$

Denote the resulting oracle estimator of the partial coefficient functions by $\widetilde{\rho}^{ij}(t), 1 \le i < j \le p$. Then the oracle estimators enjoy both estimation consistency and null region selection consistency as indicated in the following Theorem.

Theorem 1. *Under conditions (C1)-(C8), for any $1 \le i < j \le p$, the oracle estimators satisfy*

$$\begin{aligned} \left\| \widetilde{\rho}^{ij(o)} - \rho^{ij} \right\|_2 &= O_p\left(\sqrt{\frac{N_n}{nm}} + N_n^{-1} \right), \\ \sup_{t \in \mathbf{I}} \left| \widetilde{\rho}^{ij(o)}(t) - \rho^{ij}(t) \right| &= O_p\left(\frac{N_n^{3/2}}{\sqrt{nm}} + N_n^{-1} \right). \end{aligned} \tag{4.2}$$

In addition, let $\widetilde{E}^{ij} = \{t \in I, \widetilde{\rho}^{ij}(t) = 0\}$ be the corresponding null region of $\widetilde{\rho}^{ij(o)}(t)$. Then $E^{ij} \subset$

$\widetilde{E}^{ij}$, and the set $\widetilde{E}^{ij} \backslash E^{ij}$ converges to the empty set with probability approaching to 1 *as* $n \to \infty$.

Theorem 2. *Under conditions (C1)-(C8), when n is sufficiently large, the minimizer $\{\widehat{\rho}^{ij}\}_{1 \le i < j \le p}$ of the penalized likelihood function in (2.2) satisfies $\|\widehat{\rho}^{ij} - \rho^{ij}\|_2 = O_p\left(\sqrt{\frac{N_n}{nm}} + N_n^{-1}\right)$ for any $1 \le i < j \le p$.*

Theorem 3. *Under conditions (C1)-(C8), for any $1 \le i < j \le p$, let $\widehat{E}^{ij} = \{t \in I, \widehat{\rho}^{ij}(t) = 0\}$be the corresponding null region of $\widehat{\rho}^{ij}(t)$. Then $E^{ij} \subset \widehat{E}^{ij}$, and the set $\widehat{E}^{ij} \backslash E^{ij}$ converges to the empty set with probability approaching to* 1 *as* $n \to \infty$.

Theorem 2 shows that the estimator by minimizing the penalized loss function (2.2) is L_2 consistent in estimating the partial correlation functions, and Theorem 3 further shows that, with probability approaching to one, it can correctly identify zero estimators in the non-signal time regions. Therefore, the proposed method can correctly produce a locally sparse network and efficiently model the dynamic changes of network data for sufficiently large data. The proof of the Theorem is provided in the supplementary material.

Note that Theorems 2 and 3 are established under the assumption that the structure of networks changes smoothly over time (e.g, Condition C5). Therefore, the proposed spline method is developed for networks with smooth changes.

5. Simulation

In this section, we conduct simulation studies to illustrate the performance of the proposed method based on the proximal gradient method (SPG) described in Section 3. We first compare the performance of the SPG using different degrees of polynomial spline. Then the proposed approach with the best order of B-spline approximation is selected to compare with other existing approaches such as SPACE (Peng et al. 2009), the kernel-based method (Kolar et al. 2010) and the ADMM. Note that the ADMM does not apply directly in our dynamic partial correlation networks since the original ADMM is not formulated for overlapping parameters from penalty terms. Therefore we provide an adaptation of the ADMM approach to accommodate our setting. We also compare the proposed method with the time varying undirected graph (TVUG) model proposed by Zhou et al. (2010), and the varying coefficient and varying structure graphic model (VCVS) proposed by Kolar and Xing (2012). Specifically, Zhou et al. (2010) develop a kernel-based nonparametric method for estimating time-varying covariance matrices for multivariate Gaussian distributions using an l_1-regularization, and show that the TVUG model is able to obtain l_1-penalized maximum likelihood estimators at each time point as long as the covariances change smoothly over time. The VCVS model is based on the

neighborhood selection procedure (Meinshausen and Bühlmann 2006), allowing the coefficients of the precision matrix to change in a piece-wise constant fashion. That is, their model assumes that the network structures change abruptly rather than changing smoothly through incorporating both the modified fused Lasso penalty and the Lasso penalty.

We generate dynamic networks assuming that the network structures have disjointed blocks. Networks with disjointed blocks are quite common in many applications where networks are only connected within blocks, but are not associated with each other between blocks. See examples from Girvan and Newman (2002) and Valencia et al. (2009) on brain and biological functions, gene expressions, social, sports and computer network associations. In the following simulations, the number of disjointed blocks is 3. To generate the concentration matrix at time t, we first create an initial matrix $(A_t)_{p\times p}$ with three blocks as

$$\begin{pmatrix} A_t^1 & & \\ & A_t^2 & \\ & & A_t^3 \end{pmatrix},$$

where the diagonal entries for each block $A_t^k (k = 1, 2, 3)$ are all set to be one, and each off-diagonal entry of A^k is set to be $f_k(t)U$, where U follows the Bernoulli distribution with $Pr(U = 1) = \omega$. The blocks A_t^k are exchangeable since the partial correlations among nodes in networks are undirected and interchangeable. We use ω to control the number of non-zero elements in A_t^k and thus control the sparsity within each block such that the networks are sparse if ω is small. We consider moderate strengths of associations among nodes in the network, and therefore choose $\omega = 0.8$ in our settings. The functions $f_k(t), k = 1, 2, 3$ are defined as follows:

$$f_1(t) = \begin{cases} 5(t-0.5)^2 - 0.125, & \text{if} \quad 1 \le t \le 0.342 \\ 0, & \text{if} \quad 0.342 < t \le 0.658 \\ -5(t-0.5)^2 + 0.125, & \text{if} \quad 0.658 < t \le 1 \end{cases},$$

$$f_2(t) = \begin{cases} -3t + 0.9, & \text{if} \quad 0 \le t \le 0.3 \\ 0, & \text{if} \quad 0.3 < t \le 0.7 \\ 3t - 2.1, & \text{if} \quad 0.7 < t \le 1 \end{cases},$$

and

$$f_3(t) = \begin{cases} -22.5(t-0.5)^2 + 0.9, & \text{if} \quad 0.3 \le t \le 0.7 \\ 0, & \text{if} \quad \text{o.w.} \end{cases}.$$

The plots of $f_k(t)$ are provided in Figure 2. Once we construct a concentration matrix, we follow a similar strategy as in Peng et al. (2009) to ensure that the simulated covariance matrix is positive definite.

We first compare the performances of local signal selection using the linear, quadratic and cubic spline approximations in the simulation studies. Various network sizes of $p = 18, 54$ and 108, and time length $T = 50$ are considered here. The sample size is chosen as $n = 200$.

Table 1 provides the comparisons of model selection performance of the smoothing proximal gradient method (SPG) in detecting the true time-varying signals under different orders of spline approximations. Here correct-fitting (C), over-fitting (O) and under-fitting (U) are calculated as the percentages of time-points out of T equally-spaced time-points at interval $[0, 1]$ where both true-signal and non-signal points are identified correctly; true non-signal points are misclassified as signal ones; and true signal points are not selected, respectively. In addition, we also calculate sensitivity and specificity as defined by Peng et al. (2009), where sensitivity is the ratio of the number of correctly detected signals to the number of true signals; and specificity is the ratio of the number of correctly detected signals to the number of detected signals.

Table 1 indicates that the SPG with linear spline tends to select correct edges with the highest frequency, compared to the quadratic and cubic splines. When the network size increases from 18 to 108, the percentage of selecting correct associations decreases about 9.8% in the linear spline approach. When the network size is 108, the percentage of selecting correct edges based on the SPG is about 83.0% for the linear spline approach. In addition, overall sensitivity and specificity rates are best using the linear spline approach. This simulation indicates that the SPG with linear spline has the best performance in detecting the local changes of network associations, compared to the quadratic and cubic splines.

We further compare the performance of the proposed model with SPACE, the kernel-based method (KEN), the ADMM approach, the TVUG model, and the VCVS method. We compare the performance of these methods under the network sizes of 18, 54 and 108 with sample size $n = 200$ and time length $T = 50$ based on 100 simulations. Since Table 1 indicates that SPG with the linear spline outperforms the quadratic and cubic splines, we use the linear spline for the SPG in the following

comparison.

Table 2 provides the model selection performance of the SPG, ADMM, SPACE, KEN, TVUG, and VCVS under various network sizes. The SPG and ADMM have similar performance and are the best in the sense of selecting the true model with the highest frequency when the network size is 18 or 54. When the network size increases to 108, the rates of selecting the correct model for SPACE and VCVS decrease to 51.2% and 66.7%, respectively. This is probably due to the over-fitting problem. For the TVUG, the correct-fitting rate is down to 75.4%. In comparison, the SPG still has a correct-fitting rate of 83.0%. However, neither ADMM nor KEN is feasible due to the problem of high-dimensional matrix inversion for the ADMM approach and a highly intensive computing procedure for the kernel method. The ADMM requires inverting large-dimensional matrices if p is large. We tried the SparseM package in R, the Eigen package and SparseLib++ in C++ which are designed for large-dimensional matrix operations. However, when the dimension of matrices is out of the scale the package can handle, the ADMM approach becomes infeasible.

For the ADMM, the required number of iterations is $O(1/\epsilon)$ (Wang and Banerjee 2014), given a desired accuracy ϵ. For the SPG, the convergence rate is also $O(1/\epsilon)$ (Chen et al., 2012). The SPACE and the TVUG are basically a LASSO approach; the computational complexity is the same as for the quadratic programming algorithm, which is $O(n^3)$ as the worst case, where n is the sample size. For Kolar and Xing's (2012) approach, the accelerated gradient method also has a convergence rate of $O(1/\epsilon)$. For the kernel-based method, the computation complexity is due to the number of iterations, since the method only updates one parameter for each iteration. That is, if we have p nodes and m time points, the model has $p(p-1)/2 * m$-dimensional parameters, where the number of parameters increases as the number of time points increases. This leads to very intensive calculation as each iteration requires $p(p-1)/2 * m$ updates.

Table 2 also provides the average computing time per simulation run for each method. We run simulations on a cluster server running a Linux system equipped with 2.67GHz CPU and 48GB memory. The computing time increases significantly as the dimension of matrix operations increases exponentially from 10^2 to 10^5 when the network size increases from 18 to 108. The SPACE and TVUG are the fastest among all the methods. This is because the SPACE does not utilize neighboring information of the time-points observed from the same subject; and for the TVUG, the kernel-based sample covariance matrices could be pre-processed before minimization, and the covariance matrix is penalized through its determinant rather than for each element. KEN is the slowest of all since it requires updating neighborhood information for each nonparametric coefficient estimation at each

iteration. The computing time ranges from 27.46 seconds to 1.04 hours per run for the SPG algorithm, and 25.49 seconds to 15.8 minutes per run for the VCVS method. We were not able to record the time for KEN and ADMM when $p = 108$ due to infeasibility issues for these two approaches. In summary, SPG is the best among all methods above if we consider computational feasibility and correct-fitting performance.

We also compare the number of edges correctly identified by SPG, KEN, SPACE and ADMM with a moving tuning parameter. The TVUG and VCVS are not provided here since it requires two tuning parameters and makes comparison unsuitable. Figure 3 shows that the BIC reaches the minimum if the tuning parameter is selected as $\lambda = 0.145$ when the network size is 18, the sample size is 200 and the number of time-points is 50. In addition, Figure 4 indicates that both SPG and ADMM have the highest ratio of correctly identified edges over total detected edges for any given tuning parameter. For example, when the number of total detected edges equals the number of true edges (1876), the SPG and ADMM are able to identify 1444 and 1441 correct edges, respectively, whereas KEN detects 1345 correct edges, and SPACE detects only 1243 correct edges.

6. Application

In this section, we analyze a data obtained from an attention deficit hyperactivity disorder (ADHD) study. ADHD is a mental disorder found in children and adolescents, and common symptoms include being easily distracted, impulsiveness, and restlessness. To better understand how ADHD patients' brains function and react to different stimulants, we focus on identifying associations and interactions among different regions of interest (ROI) of the brain. One distinct feature of ADHD patients is to have high variability of brain function over time; therefore, it is scientifically important to identify the dynamic changes of association among different regions of interest of the brain to locate the ADHD pathology.

The ADHD-200 samples contain fMRI data which are repeatedly measured over time. The data are downloaded from `http://www.nitrc.org/frs/?group_id=383|`, which contains resting-state fMRI (rs-fMRI) of 78 patients (mean age=9.0 and s.d.=1.12) from the Oregon Health & Science University with 116 regions of interest measured over 74 time points. The software for automated anatomical labeling was used to label macroscopic brain structures which categorize the brain into 116 regions of interest (`http://neuro.imm.dtu.dk/wiki/Automated_Anatomical_Labeling|`). The patients were instructed to stay still, keep their eyes open and focus on a standard fixation cross in the center of the display. Participants were scanned after a minimum washout of short-acting stimulant medications. The temporal-resolution of fMRI data is 2500

ms.

We apply only the SPG and SPACE methods to this data, since the ADMM and KEN approaches are not able to handle the network size of 116. The number of connections among ROIs at each time point is shown in Table 3. Note that SPACE identifies more than 2000 connections at most of the time-points, in contrast to the SPG method which identifies at most 78 connections at each time point. The over-identifying problem of SPACE makes it difficult to select any useful connections. In the following, we provide data analysis and graphical illustration based on the SPG only.

Figure 5 illustrates the associations and connections of 116 regions of interest formulated as a network at time points $t = 1, 10, 20, 50, 60$ and 74. Note that each region of interest in the brain is represented as nodes or vertices with either green or pink color, and the associations among nodes are connected with blue lines. The color pink of a node represents five or more associations with other regions of interest, and the color green of a node indicates less than five associations with other regions of interest.

We are able to identify the dynamic changes of associations among the 116 regions of interest over time. Specifically, the ADHD patients experience three distinct periods of brain activities during the test. The number of connections at each time point is shown in Table 3. At the beginning of the test, the ADHD patients' brains are active. However, when the test proceeds further, the ADHD patients' brains are mostly in a resting state, since there are only a few connections among the 116 regions of interest, with most of the regions of interest containing less than 36 connections. This is possibly due to the fact that patients are less disturbed in the middle of the experiment, since there is actually no stimulus imposed on their brains. In the later stage of the test when $t > 57$, patients' brains again have more connections among regions of interest, as patients might anticipate something happening by the end of the experiment. These phenomena are also indicated in Figure 5, showing that there are more associations among regions of interest between $t = 1$ and $t = 10$, and $t = 60$ and $t = 74$, but fewer brain activities between $t = 20$ and $t = 50$.

Table 4 confirms our findings and indicates that there are few associations between $t = 20$ and $t = 55$, with only 2 vertices having three or more connections during this period. However, between time points $t = 1$ and 19, there are 15 vertices containing three or more connections among regions of interest, and between $t = 56$ and 74, there are 14 vertices having three or more connections. The corresponding names of those ROIs with three or more connections and their gray levels are provided in Table 5 (gray level is defined as the volume of gray matter in a ROI, and gray matter distinguished from white matter consists of cell bodies, neuropil, glial cells and capillaries). These findings could

be helpful in studying ADHD patients' brain function over time, even without any stimulation.

Compared to task-based fMRI experiments, results from resting-state fMRI studies can be more easily synthesized as they investigate the differences for the ADHD patients' regions of interest connected in the absence of task. Fox and Greicius (2010) and Greicius (2008) studied the connections between any two regions of interest, and used two-sample t-tests to infer whether the average strength of connection between two regions of interest is significantly different between ADHD and healthy patients. Dickstein et al. (2006) also found that there were several regions of interest consistently under-activated among patients with ADHD. These include portions of the frontal lobe: anterior cingulate cortex (ACC) (regions 31 and 32 in AAL), dorsolateral prefrontal cortex (DLPFC), and inferior prefrontal cortex (11-16, AAL), along with portions of the basal ganglia, thalamus, and parietal cortices. Hart et al. (2013) discovered that portions of the frontal lobe (the inferior frontal cortex, ACC, and supplemental motor area), basal ganglia and thalamus are under-activated in response to inhibition tasks among ADHD patients. Furthermore, patients with ADHD showed under-activation in the DLPFC, parietal areas, basal ganglia and thalamus in response to attention tasks. In Figure 5, we highlighted the nodes in our network graphs. Nodes 11-16 and 31, 32 are not active, except that node 32 becomes active at the end of the test (with 4 connections). So, in general, our analysis results are consistent with the findings in the existing literature, as mentioned above.

Figure 6 describes the changes of associations among three ROIs (right middle frontal gyrus, right gyrus rectus and right angular gyrus) at $t = 1, 20$, and 60. The regions of interest are highlighted as green if they are associated with each other at certain time points. Figure 7(a) illustrates the locations of certain ROIs in the brain using an automated anatomical labeling (AAL) software package. Here different ROIs are marked as different colors. Note that most of the ROIs have counterparts located on the opposite side of the brain, and are marked as the same color. For example, the cyan blue color is used for both Temporal_Mid_L and Temporal_Mid_R in Figure 7(a). However, these counterpart ROIs are not necessarily associated with each other. Figure 7(a) shows 50 out of the 116 ROIs, and Figure 7(b) provides a partial network of the ROIs to illustrate the associations based on the selected 15 ROIs. The partial network is quite sparse. For better visualization of the associated network, Figure 7(c) also provides the associated names of the 15 selected ROIs.

In addition, we also provide an animated video in the file "ADHD.mp4" to illustrate the dynamic changes for 116 regions of interest of the brain over 74 time points. The colors of the nodes in the video ranges from red to purple, blue and green, which reflects the level of connections with other ROI over the entire time period. The red nodes are the most active ROIs with the number of connections

ranging from 30 to 36; the purple nodes have a number of connections from 18 to 29; while the blue and green nodes have moderate to few associations with other ROIs of the brain, ranging between 8 to 17 and 0 to 7, respectively.

7. Discussion

The time-varying network model is powerful for identifying time-evolving associations for brain and biological functions, gene networks, social networks and environmental networks over time. In this paper, we develop a local varying-coefficient model to effectively quantify and detect dynamic changes in network associations and interactions. One distinctive feature of the proposed approach is that we are able to incorporate local features of a varying-coefficient function, and provide local-signal detection and estimation simultaneously for time-varying network data.

We propose a piecewise penalized loss function such that the coefficients associated with the varying-coefficient model at the local region are shrunk to zero if the magnitude of the grouped coefficients is sufficiently small. This has significant advantages over the traditional varying-coefficient model selection approach without incorporating local features, especially for time-varying network data, since the network associations could be quite volatile over time, and local-region estimation and signal detection are of more scientific interest than global-feature selection. Our simulation studies and data application to the ADHD study indicate that the proposed method is quite effective at capturing the local features of the time-varying network data.

However, it is quite computationally challenging to develop highly computationally intensive algorithms in order to achieve sparsity properties in estimation and signal detection at local time intervals. The group penalization strategy involves overlapping parameters among different groups, which makes the optimization process extremely challenging when the network size is large. To overcome these difficulties, we develop a smoothing proximal gradient method, which does not require inverting the large-dimensional matrix. The developed algorithm has significant computational advantages in increasing computational speed and efficiency. Most importantly, the proposed smoothing proximal gradient algorithm is able to analyze a relatively large size of network data within a reasonable time frame. We also compare out method to the ADMM and kernel-based algorithms which require inverting a large-dimensional matrix, and therefore cannot feasibly estimate large size network data.

Theoretically, we show that the proposed method achieves model selection consistency in local regions, and provides a uniform rate of convergence for local-signal coefficient estimators. Scientifically, it is important to detect dynamic changes in networks, as identifying the associations of biological functionalities over time can help us to better understand the mechanisms of network change.

The proposed method is developed for networks with a fixed dimension. For a high-dimensional network, we suggest to first use some screen methods to bring the dimensionality down. For example, one can use a global selection method similar to the ones in Xue (2009), Xue and Qu (2012) to delete the pairs of variables that not associated/connected in the entire region. Then for the pairs that are associated, one then can apply the proposed method to locate the time region where this association might change.

In this paper, the longitudinal dependence structure over time is not taken into account for estimation, although one can incorporate such dependent structure using either the generalized estimation equation (Liang and Zeger 1986) or the quadratic inference function approach as in Xue et al. (2010) and Wang et al. (2014). However, incorporating the dependence structure does not effect the convergence rate as in Section 4, but will effect estimation efficiency. This might be worthy of future research.

Supplementary Materials

The document includes detailed proofs of main Theorems and necessary Lemmas.

Acknowledgements

Xue's research was supported by the Simons Foundation (F0782A). Qu's research was supported by the National Science Foundation (DMS-1308227, DMS-1415308 and DMS-1613190).

References

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

[2] Chen, Z. and Leng, C. (2016). Dynamic covariance models. *Journal of the American Statistical Association*, **111**, 1196-1207.

[3] Chen, X., Lin, Q., Kim, S., Carbonell, J. G. and Xing, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression. *Ann. Applied Statist.* **6**, 719-752.

[4] Cheng, M-Y , Honda, T., and Zhang, J-T (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **111**, 1209-1221.

[5] Chung, F. R. K. (1997). Spectral graph theory. *CBMS Regional Conference Series in Mathematics*, No. 92.

[6] Dickstein, S. G., Bannon, K., Castellanos, F. X. and Milham, M. P. (2006). The neural correlates of attention deficit hyperactivity disorder: An ALE meta-analysis. *J. Child Psychol Psychiatry* **47**, 1051-1062.

[7] Fan, J. and Gijbels, I. (1996) *Local polynomial modelling and its applications.* Chapman and Hall, London.

[8] Fox, M.D. and Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Front Syst Neurosci* **4**, 19.

[9] Friedman, J. H., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

[10] Friston, K. J., Harrison, L. and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* **19**, 1273-1302.

[11] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl Acad. Sci.* **99**, 7821-7826.

[12] Greicius M (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Curr Opin Neurol* **21**, 424-430.

[13] Guo, J., Levina, L., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1-15.

[14] Hart, H., Radua, J., Nakao, T., Mataix-Cols, D. and Rubia, K. (2013). Meta-analysis of functional magnetic resonance imaging studies of inhibition and attention in attention deficit/hyperactivity disorder: Exploring task-specific, stimulant medication, and age effects. *JAMA Psychiatry* **70**, 185-198.

[15] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

[16] Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242-272.

[17] Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficent models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**, 111-128.

[18] Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14**, 763-788.

[19] Huang, J. Z., Zhang, L. and Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scandinavian Journal of Statistics* **34**, 451-477.

[20] Jacob, L., Obozinski, G. and Vert, J. P. (2009). Group lasso with overlap and graph lasso. *In Proceedings of the International Conference on Machine Learning*.

[21] Jenatton, R., Audibert, J.-Y. and Bach, F. (2011). Structured variable selection with sparsity inducing norms. *J. Mach. Learn. Res.* **12**, 2777-2824.

[22] Kim, W. H., Adluru, N., Chung, M. K., Charchut, S., GadElkarim, J. J., Altshuler, L., Moody, T., Kumar, A., Singh, V. and Leow, A. D. (2013). Multi-resolutional brain network filtering and analysis via wavelets on non-Euclidean space. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, 643-651.

[23] Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and models.* New York, Springer.

[24] Kolar, M., Parikh, A. and Xing, E. P. (2010). On sparse nonparametric conditional covariance selection. *The 27th International Conference on Machine Learning*.

[25] Kolar, M., Song, L. and Xing, E. P. (2009). Sparsistent learning of varying-coefficient models with structural changes. *Advances in Neural Information Processing Systems* **23**, 1006-1014.

[26] Kolar, M. and Xing, E. P. (2009). Sparsistent estimation of time-varying discrete Markov random fields. *arXiv preprint arXiv:0907.2337.*

[27] Kolar, M. and Xing, E. P. (2011). On time varying undirected graphs. *J. Mach. Learn. Res.* **15**, 407-415.

[28] Kolar, M. and Xing, E. P. (2012). Estimating networks with jumps. *Electronic Journal of Statistics* **6**, 2069-2106.

[29] Lebre, S., Becq, J., Devaux, F., Stumpf, M. P. and Lelandais, G. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology* **4**, 130-145.

[30] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[31] Lee, H., Lee, D. S., Kang, H., Kim, B. N. and Chung, M. K. (2011). Sparse brain network recovery under compressed sensing. *IEEE Transactions on Medical Imaging* **30**, 1154-1165.

[32] Leonardi, N. and Van De Ville, D. (2011). Wavelet frames on graphs defined by fMRI functional connectivity. 2011 IEEE International Symposium, 2136-2139.

[33] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the LASSO. *Ann. Statist.* **34**, 1436-1462.

[34] Obozinski, G., Jacob, L. and Vert, G. (2011). Group lasso with overlaps: The latent group lasso approach. *arXiv preprint arXiv:1110.0413.*

[35] Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104**, 735-746.

[36] Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* **62**, 379-391.

[37] Shen, X., Huang, H. and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika* **99**, 899-914.

[38] Shojaie, A. and Michailidis, G. (2010). Network enrichment analysis in complex experiments. *Stat. Appl. Genet. Mol. Biol.* **9**, Art. 22.

[39] Song, L., Kolar, M. and Xing, E. P. (2009). KELLER: Estimating time-evolving interactions between genes. *Bioinformatics* **25**, 128-136.

[40] Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E., Breakspear, M. and Friston, K. J. (2008). Nonlinear dynamic causal models for fMRI. *Neuroimage* **42**, 649-662.

[41] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. Ser. B.* **58**, 267-288.

[42] Tibshirani, R., Sauders, M., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J R Statist Soc B.* **67**, 91-108.

[43] Valencia, M., Pastor, M. A., Fernadez-Seara, M. A., Artieda, J., Martinerie, J. and Chavez, M. (2009). Complex modular structure of large-scale brain networks. *Chaos* **19**, 023-119.

[44] Wang, H. and Banerjee, A. (2014). Bregman alternating direction method of multipliers. *Advances in Neural Information Processing System*, 2816-2824.

[45] Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

[46] Wang, L., Xue, L., Qu, A. and Liang, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *Ann. Statist.* **42**, 592-624.

[47] Wang, R., Lin, P. and Wu, Y. (2015). Exploring dynamic temporal-topological structure of brain network within ADHD. *In Advances in Cognitive Neurodynamics (IV)*, 643-651.

[48] Wee, C. Y., Yang, S., Yap, P. T., Shen, D. and Alzheimer's Disease Neuroimaging Initiative. (2016). Sparse temporally dynamic resting-state functional connectivity networks for early MCI identification. *Brain imaging and behavior* **10**, 342-356.

[49] Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high-dimensional varying coefficient models. *Statist. Sinica.* **21**, 1515-1540.

[50] Wee, C.-Y., Yap, P.-T., Zhang, D., Wang, L. and Shen, D. (2012). Constrained sparse functional connectivity networks for MCI classification. *Medical Image Computing and Computer-assisted Intervention - MICCAI* **15**, 212-219.

[51] Xue, L. (2009). Variable selection in additive models. *Statist. Sinica.* **19**, 1281-1296.

[52] Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Mach. Learn. Res.* **13**, 1973-1998.

[53] Xue, L., Qu, A. and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *J. Amer. Statist. Assoc.* **105**, 1518-1530.

[54] Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. *Statistica Sinica*, 1423-1446.

[55] Zhou, J., Wang, N.Y. and Wang, N. (2013). Functional linear model with zero-value coefficient function at subregion. *Statist. Sinica.* **23**, 25-50.

[56] Zhou, S., Lafferty, J. and Wasserman, L. (2010). Time-varying undirected graphs. *Machine Learning Journal.* **80**, 295-319.

[57] Zhu, Y., Shen, X. and Pan, W. (2013). Simultaneous grouping pursuit and feature selection in regression over an undirected graph. *J. Amer. Statist. Assoc.* **108**, 713-725.

[58] Zhu, Y., Shen, X. and Pan, W. (2014). Structural pursuit over multiple undirected graphs. *J. Amer. Statist. Assoc.* **109**, 1683-1696.

[59] Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Statistics, Oregon State University

E-mail: xuel@science.oregonstate.edu

Merck, Kenilworth, NJ

E-mail: xinxin.shu@merck.com

Department of Statistics, University of Illinois at Urbana-Champaign

E-mail: anniequ@illinois.edu

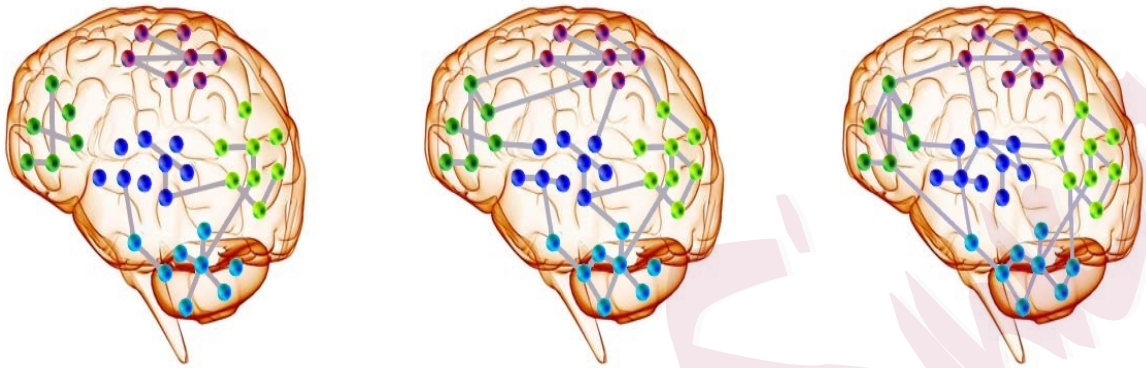
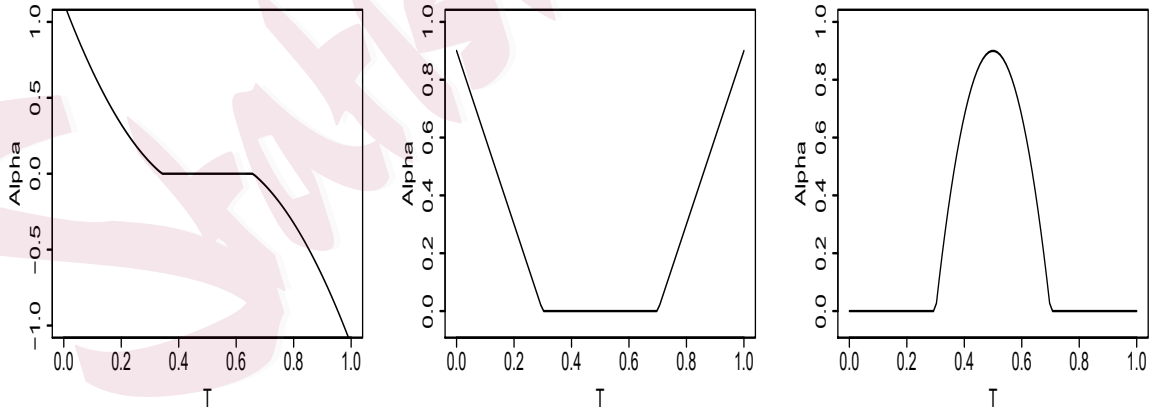


Figure 1: Changes of associations among different sites of a brain over three time-points



(a) $f_1(t)$ (b) $f_2(t)$ (c) $f_3(t)$

Figure 2: The function $f(t)$ at time interval $t \in [0, 1]$

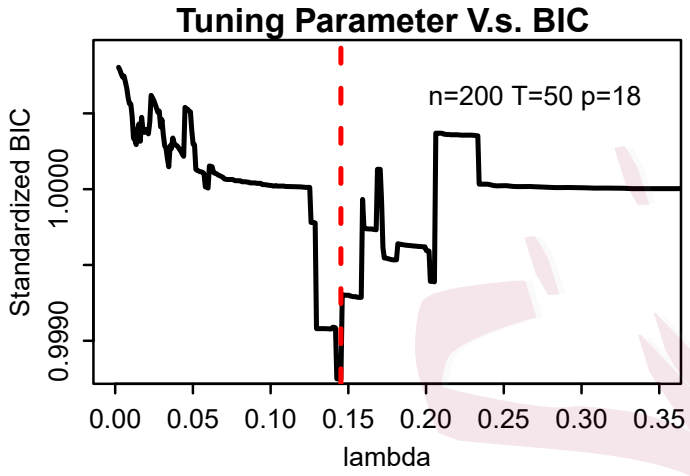


Figure 3: The plot of moving tuning parameter versus the BIC for the SPG algorithm when $n = 200$, $T = 50$ and $p = 18$.

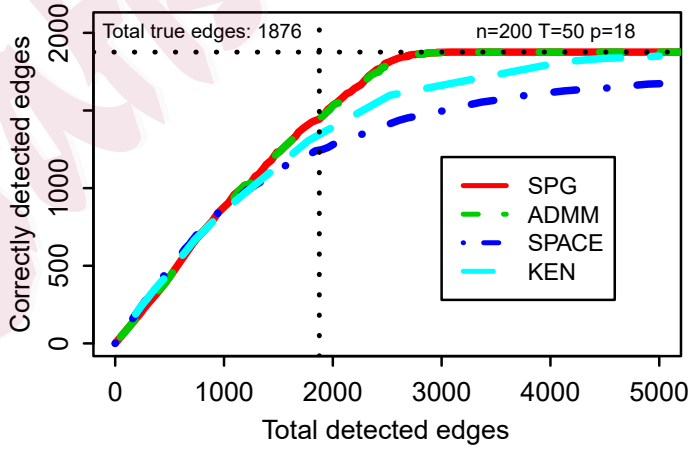
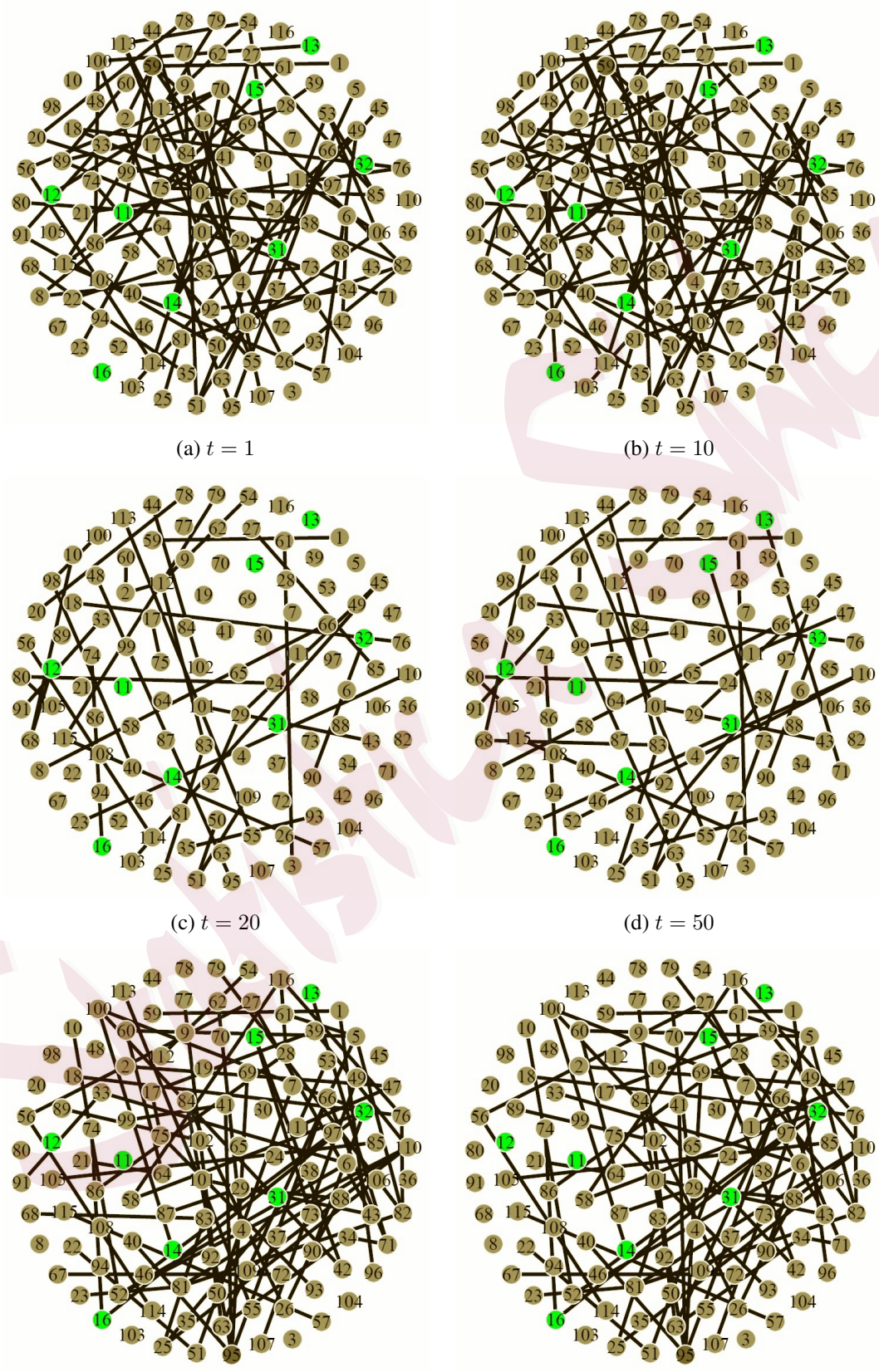


Figure 4: Correctly detected edges versus total detected edges using the four methods.



(a) $t = 1$ (b) $t = 10$

(c) $t = 20$ (d) $t = 50$

(e) $t = 60$ (f) $t = 74$

Figure 5: Estimation of brain networks of ADHD-200 data at time-points $t = 1, 10, 20, 50, 60$ and 74.

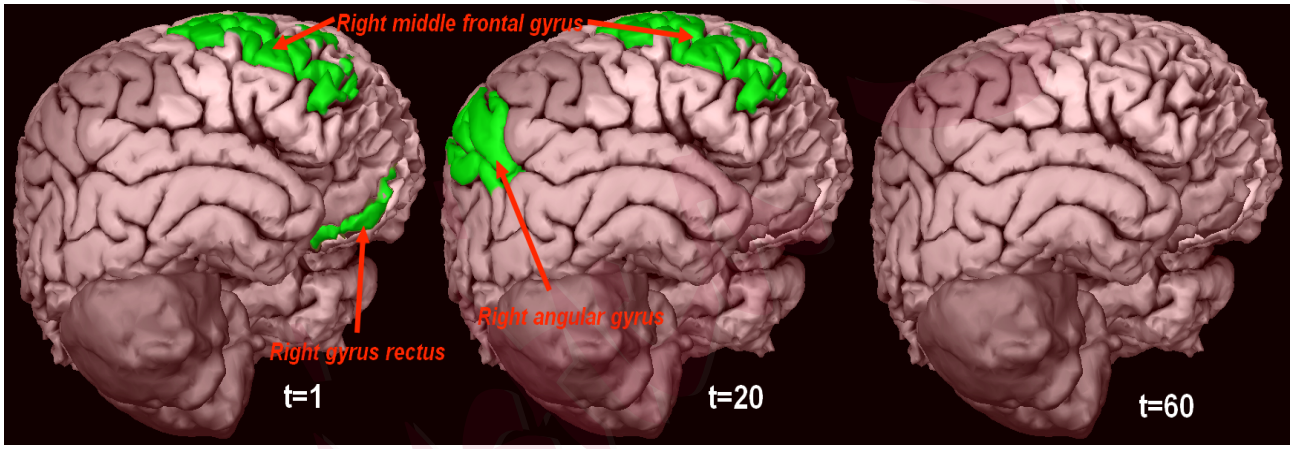
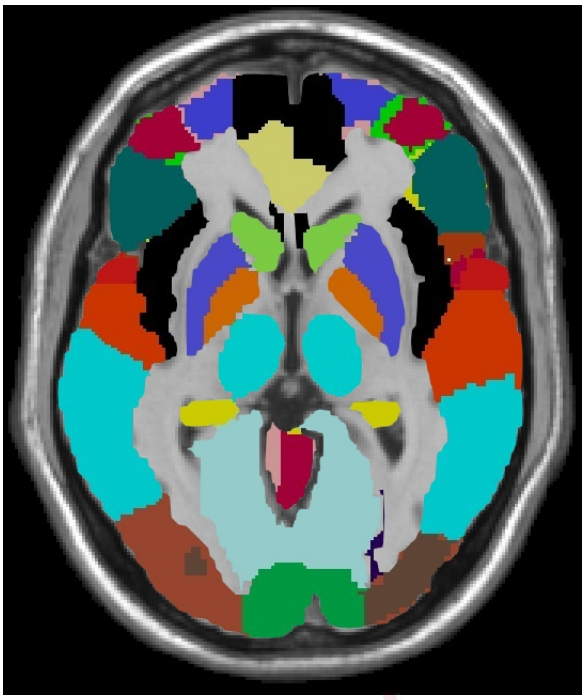
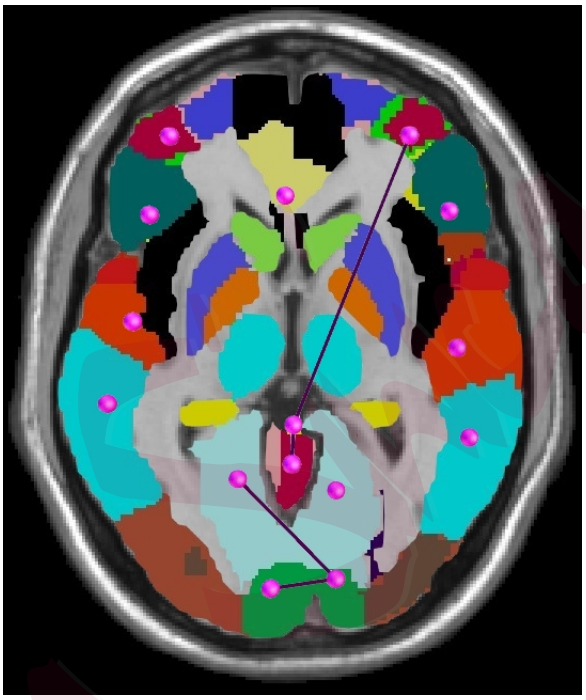


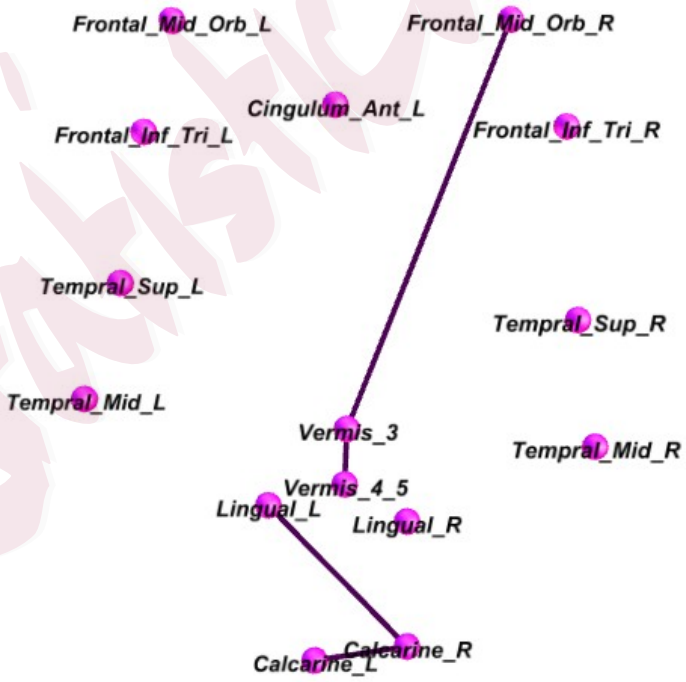
Figure 6: Changes of associations among right middle frontal gyrus, right gyrus rectus, and right angular gyrus over three time-points



(a) AAL ROIs' locations in brain



(b) The partial network of ROIs



(c) ROIs' names in the partial network

Figure 7: Illustration of AAL ROIs in the brain and its networks

Table 1: Model selection performance of the smoothing proximal gradient method (SPG) for three-block disjointed networks with the number of time-points $T = 50$ and sample size 200 based on 100 simulation runs.

	Network size	C	O	U	Sensitivity	Specificity	Time
Linear	p=18	0.920	0.056	0.024	0.802	0.967	27.46
	p=54	0.859	0.071	0.070	0.679	0.910	467.71
	p=108	0.830	0.023	0.147	0.772	0.836	3726.48
Quadratic	p=18	0.887	0.063	0.050	0.760	0.932	41.87
	p=54	0.838	0.073	0.089	0.642	0.888	670.89
	p=108	0.799	0.088	0.113	0.560	0.859	7510.36
Cubic	p=18	0.860	0.091	0.049	0.688	0.931	60.13
	p=54	0.791	0.099	0.110	0.526	0.861	1192.30
	p=108	0.764	0.113	0.123	0.474	0.843	14102.38

Table 2: Model selection performance of SPG, ADMM, SPACE, KEN and VCVS for three-block disjointed networks with the number of time-points $T = 50$ and sample size 200 based on 100 simulation runs.

Network size	Methods	C	O	U	Sensitivity	Specificity	Time per run (seconds)
p=18	SPG	0.920	0.056	0.024	0.802	0.967	27.46
	ADMM	0.920	0.055	0.025	0.804	0.965	10.53
	SPACE	0.907	0.082	0.011	0.745	0.984	1.33
	KEN	0.909	0.065	0.026	0.775	0.963	109.35
	TVUG	0.880	0.079	0.041	0.726	0.942	2.03
	VCVS	0.901	0.052	0.047	0.796	0.937	25.49
p=54	SPG	0.859	0.071	0.070	0.679	0.910	467.71
	ADMM	0.860	0.068	0.072	0.685	0.908	286.87
	SPACE	0.691	0.220	0.089	0.373	0.863	36.39
	KEN	0.786	0.123	0.091	0.512	0.878	14328.74
	TVUG	0.820	0.096	0.084	0.586	0.891	26.79
	VCVS	0.748	0.127	0.124	0.430	0.840	123.94
p=108	SPG	0.830	0.023	0.147	0.772	0.836	3726.48
	ADMM	NA	NA	NA	NA	NA	NA
	SPACE	0.512	0.418	0.070	0.271	0.836	349.98
	KEN	NA	NA	NA	NA	NA	NA
	TVUG	0.754	0.136	0.110	0.458	0.853	383.76
	VCVS	0.667	0.220	0.113	0.337	0.831	944.03

Table 3: Number of associations identified by SPG and SPACE from time-points 1 to 74.

Method	Number of associations from 1 to74
SPG	70 77 77 77 76 77 77 76 77 77 77 77 77 77 78 77 77 77 77 35 36 35 35 35 35 35 35 35 35 35 35 35 35 36 35 35 35 34 34 34 34 34 34 34 34 34 34 34 34 34 35 34 34 34 34 76 76 76 75 76 76 76 76 76 77 76 76 76 76 76 76 76 66
SPACE	3024 3102 3257 2059 2691 2839 3278 2962 3111 3080 2926 2946 2833 3079 3171 3156 3067 2932 3129 2955 2934 3025 1998 3076 3130 3278 3230 2786 3176 2828 2979 2981 3057 3045 2695 3070 2665 3120 3090 2916 3054 2982 2670 3038 2836 2969 3006 3154 2756 3056 3179 3024 2975 2974 3067 3273 1956 3157 2707 3132 3115 2948 2799 2967 3028 3059 2969 3165 3089 3039 3109 2950 3103 2779

Table 4: ROIs with 5 or more associations identified by SPG from time-points 1 to 74.

Time(t)	ROIs with 3 or more associations	Total
1-19	24 38 51 53 54 59 70 75 82 85 89 100 106 113 115	15
20-55	83 112	2
58-74	5 25 32 52 63 71 76 81 82 90 95 100 110 116	14

Table 5: Name list of ROIs with 3 or more associations identified by SPG

Number	Name	Gray level
5	Frontal_Sup_Orb_L	2111
24	Frontal_Sup_Medial_R	2602
25	Frontal_Mid_Orb_L	2611
32	Cingulum_Ant_R	4002
38	Hippocampus_R	4102
51	Occipital_Mid_L	5201
52	Occipital_Mid_R	5202
54	Occipital_Inf_R	5302
59	Parietal_Sup_L	6101
63	SupraMarginal_L	6211
70	Paracentral_Lobule_R	6402
71	Caudate_L	7001
75	Pallidum_L	7021
76	Pallidum_R	7022
82	Temporal_Sup_R	8112
83	Temporal_Pole_Sup_L	8121
85	Temporal_Mid_L	8201
89	Temporal_Inf_L	8301
90	Temporal_Inf_R	8302
95	Cerebelum_3_L	9021
100	Cerebelum_6_R	9042
106	Cerebelum_9_R	9072
110	Vermis_3	9110
112	Vermis_6	9130
113	Vermis_7	9140
115	Vermis_9	9160
116	Vermis_10	9170