Time-Varying Dynamic Bayesian Network Learning for an fMRI Study of Emotion Processing

Lizhe Sun, Aiying Zhang and Faming Liang *

Abstract

This paper presents a novel method for learning time-varying dynamic Bayesian networks. The proposed method breaks down the dynamic Bayesian network learning problem into a sequence of regression inference problems and tackles each problem using the Markov neighborhood regression technique. Notably, the method demonstrates scalability concerning data dimensionality, accommodates time-varying network structure, and naturally handles multi-subject data. The proposed method exhibits consistency and offers superior performance compared to existing methods in terms of estimation accuracy and computational efficiency, as supported by extensive numerical experiments. To showcase its effectiveness, we apply the proposed method to an fMRI study investigating the effective connectivity among various regions of interest (ROIs) during an emotion-processing task. Our findings reveal the pivotal role of the subcortical-cerebellum in emotion processing.

Keywords: Dynamic Bayesian Network, Sparse Graphical Model, Markov Neighborhood Regression, Variable Selection, Brain Connectivity

*To whom correspondence should be addressed: F. Liang. Liang is Distinguished Professor (email: fmliang@purdue.edu), Department of Statistics, Purdue University, West Lafayette, IN 47907. Sun is Postdoc, Beijing International Center for Mathematical Research, Peking University and Department of Statistics, Purdue University. Zhang is Assistant Professor, School of Data Science, University of Virginia.

1 Introduction

Functional magnetic resonance imaging (fMRI) is a neuroimaging technique that provides a noninvasive measure of neuronal activity in the human brain by evaluating changes in blood oxygenation levels. When a particular stimulus activates a brain area, the local oxygen consumption increases rapidly, and oxygen-rich blood flows to that area, leading to an increase in oxyhemoglobin and a decrease in deoxyhemoglobin in the activated region. The blood oxygenation-level dependent (BOLD) signal measures the difference between the levels of oxyhemoglobin and deoxyhemoglobin in local blood flow. Typically, fMRI experiments involve collecting massive BOLD time series data from spatially distinct brain locations.[1]

Learning brain connectivity networks[2] is an important task in fMRI studies as it can help people understand how different brain regions collaborate to address specific cognitive processes.[3] This task can be addressed statistically by learning a dynamic graphical model, where each node represents a region of interest (ROI) and each edge represents the brain connectivity among the ROIs. From fMRI data, two types of brain connectivity can be inferred: undirected functional connectivity and effective connectivity.[4]

Undirected functional connectivity refers to the temporal correlation or dependence among ROIs, revealing general patterns of communication among brain regions.[5;6] In the literature, a variety of methods have been developed to infer undirected dynamic functional connectivity, such as sliding window analysis,[7] hidden Markov modeling,[8] and joint estimation of multiple graphical models.[9–13] Unfortunately, these analyses only provide information on the functional connectivity between brain regions while losing the direction of information flow between them.

Effective connectivity analysis reveals the direction of the information flow during brain activity and can determine whether one brain region communicates downstream to another brain region or vice versa. Typically, effective connectivity varies over time during brain activities[14]. The vector auto-regression (VAR) model[4;15] and dynamic Bayesian network[16;17] have often been used for inferring effective connectivity in fMRI studies. However, these methods are often less scalable with respect to the data dimension (i.e., the number of ROIs) due to the computational complexity involved in large-scale matrix inversion[4;18] or intensive

MCMC simulations[16]. As a result, they can only handle small brain connectivity networks. Additionally, the use of these methods can be limited by the stationarity assumption for the time series and the availability of subject-level data. In particular, these methods often assume that the underlying model parameters remain constant within a task[4;19], and the analysis often focuses on a single subject[8;20]. How to efficiently perform effective connectivity analysis for time-varying, high-dimensional, and multi-subject fMRI data has posed a great challenge to the existing methods.

In this paper, we address the challenge by developing a new dynamic Bayesian network learning method that is scalable with respect to the data dimension, allows for time-varying underlying models, and naturally handles multi-subject data. The proposed method breaks down the dynamic Bayesian network learning problem into a series of high-dimensional statistical inference problems. At each time point and for each ROI, the directly dependent ROIs are identified using the Markov neighborhood regression (MNR) method[21], and the Markov neighborhoods required for each of the regression tasks involved are identified using a Joint multiple Gaussian Graphical Models (JGGM) estimation method.[22] Since each regression task is based on observations from two neighboring time points only, the proposed method allows the underlying dynamic system to vary significantly over time. The MNR breaks down the high-dimensional inference problems into a series of low-dimensional inference problems based on the graph theory of Markov blankets, thereby avoiding high-dimensional matrix inversion. The JGGM method breaks down the high-dimensional graphical model construction problems into a series of low-dimensional conditional independence tests based on an equivalent measure of partial correlation coefficient[23]. The appealing dimension reduction feature of both the MNR and JGGM methods enables the proposed method to be scalable with respect to the data dimension. The proposed method is illustrated through an fMRI study of emotion processing[24], and our results suggest that the subcortical-cerebellum plays a crucial role in emotion processing. Additionally, we find that there is increased activity in the inter-modular connectivity among the subcortical-cerebellum, motor, visual II, and visual association modules during the period of emotion processing.

The remaining part of this paper is organized as follows. Section 2 gives a brief introduction to the MNR method. Section 3 describes the time-varying dynamic Bayesian

network model for effective connectivity analysis. Section 4 describes the proposed method for learning time-varying dynamic Bayesian networks. Section 5 provides a brief introduction to task-based fMRI data and then assesses the performance of the proposed method via simulated fMRI data. Section 6 learns dynamic Bayesian networks for a real fMRI dataset of emotion processing. Section 7 concludes the paper with a brief discussion.

2 Markov Neighborhood Regression

The MNR method was first proposed by Liang et al[21] for high-dimensional inference and constructing the causal graph structure around the response variable Y in high-dimensional regression. Compared to other high-dimensional inference methods, such as desparsified Lasso,[25–27] the MNR successfully decomposes the high-dimensional inference problem into a series of low-dimensional inference problems and it produces more accurate estimates for confidence intervals and p-values. A brief review of the MNR method is provided below.

Suppose that a set of n independent samples $D_n = \{(Y^{(i)}, \boldsymbol{X}^{(i)})_{i=1}^n\}$ have been collected from the linear regression with a random design:

$$Y = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p + \epsilon, \tag{1}$$

where ϵ follows the normal distribution $N(0, \sigma^2)$, and the covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$ follows a multivariate normal distribution $N_p(\boldsymbol{0}, \Sigma)$. Let $\boldsymbol{\varpi}^* = \{X_j : \beta_j \neq 0\}$ denote the set of true variables of the model (1). Suppose that $\boldsymbol{X}$ is represented by a Gaussian graphical model (GGM) denoted by $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{1, 2, \ldots, p\}$ is the set of p vertices, and $\boldsymbol{E} = (e_{ij})$ represents the adjacency matrix. Here, $e_{ij} = 1$ indicates a link between nodes i and j, while $e_{ij} = 0$ indicates the absence of a link.

Let $\boldsymbol{X}_A = \{X_k : k \in A\}$ denote a set of features indexed by $A \subset \boldsymbol{V}$. Let $\xi_j = \{k : e_{jk} = 1\}$ denote the neighboring set of X_j in $\boldsymbol{G}$. It follows from the Markov property of the GGM that $X_j \perp X_i | \boldsymbol{X}_{\xi_j}$ for any $i \in \boldsymbol{V} \setminus \xi_j$, where $X_a \perp X_b | \boldsymbol{X}_{\boldsymbol{c}}$ denotes the conditional independence of X_a and X_b given $\boldsymbol{X}_{\boldsymbol{c}}$. The set ξ_j is called the minimum Markov neighborhood of X_j in $\boldsymbol{G}$. The minimum Markov neighborhood is also termed as the Markov blanket in graphical theory. Any subset $\tilde{D}_j$ is a Markov neighborhood of X_j if $\xi_j \subseteq \tilde{D}_j \subseteq \boldsymbol{V} \setminus \{j\}$. The conditional

independence relationships implied by the Markov neighborhood form the basis for statistical inference in the MNR method.

Without loss of generality, we let $\tilde{D}_1 = \{2, \ldots, d\} \supseteq \xi_1 \cup \boldsymbol{\varpi}^* \setminus \{1\}$ denote a Markov neighborhood of X_1, let Σ_d denote the covariance matrix of $\{X_1\} \cup \boldsymbol{X}_{\tilde{D}_1}$, and partition $\Theta = \Sigma^{-1}$ as

$$\Theta = \begin{bmatrix} \Theta_d & \Theta_{d,p-d} \\ \Theta_{p-d,d} & \Theta_{p-d} \end{bmatrix}. \tag{2}$$

Following from the well known property of the GGM,[28] for any variables X_i and X_j,

$$X_j \perp X_i | X_{\boldsymbol{V}\setminus\{i,j\}} \Longleftrightarrow \theta_{ij} = 0, \tag{3}$$

where θ_{ij} denotes the (i,j)-th entry of Θ. Therefore, the first row of $\Theta_{d,p-d}$ and the first column of $\Theta_{p-d,d}$ in (2) are exactly zero, as $X_1 \perp \boldsymbol{X}_{\boldsymbol{V}\setminus(\{1\}\cup\tilde{D}_1)} | \boldsymbol{X}_{\tilde{D}_1}$ holds. Inverting Θ, we have $\Sigma_d = (\Theta_d - \Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d})^{-1}$, which is equal to the top $d \times d$-submatrix of Σ. Therefore,

$$\Sigma_d^{-1} = \Theta_d - \Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d}. \tag{4}$$

Since the first row of $\Theta_{d,p-d}$ and the first column of $\Theta_{p-d,d}$ are exactly zero, the $(1,1)$-th element of $\Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d}$ is exactly zero. Therefore, the $(1,1)$-th entry of Θ_d (and Θ) equals to the $(1,1)$-th entry of Σ_d^{-1}. This suggests that the statistical inference for β_1 can be made based on the subset regression:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_d\beta_d + \epsilon. \tag{5}$$

Since $\tilde{D}_1$ forms a Markov neighborhood of X_1 in the GGM, the method is called *Markov neighborhood regression* (MNR).

Let $\hat{\xi}_j$ denote an estimate of ξ_j, let $\hat{\boldsymbol{\varpi}}^*$ denote an estimate of $\boldsymbol{\varpi}^*$, and let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{\boldsymbol{\varpi}}^*$. Liang et al[21] established the validity of the MNR method under the assumptions:

$$\hat{\boldsymbol{\varpi}}^* \supseteq \boldsymbol{\varpi}^*, \tag{6}$$

$$\hat{\xi}_j \supseteq \xi_j, \ \forall j \in \{1, 2, \ldots, p\}, \tag{7}$$

$$|D_j| = |\{j\} \cup \hat{\xi}_j \cup \hat{\boldsymbol{\varpi}}^*| = o(\sqrt{n}). \tag{8}$$

For the validity of these assumptions, please refer to Section 2.1 of Liang et al.[21] For each $j \in \boldsymbol{V}$, if the conditions (6)-(8) are satisfied, then $\sqrt{n}(\hat{\beta}_j - \beta_j) \sim N(0, \sigma^2\theta_{jj})$ as shown in

Liang et al[21], where θ_{jj} is the (j,j)-th entry of the precision matrix Θ, and $\hat{\beta}_j$ is the ordinary least square (OLS) estimate of β_j obtained by regressing Y on $\boldsymbol{X}_{D_j}$.

For the case that n is finite, one can use the student-$t(n-|D_j|-1)$ distribution to approximate the distribution of $\sqrt{n}\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}_n^2\hat{\theta}_{jj}}}$; that is, *the estimate, p-value (for the test of the hypothesis $\beta_j=0$ based on $\hat{\beta}_j$) and confidence interval of β_j can be calculated from a subset regression as in conventional low-dimensional multiple linear regression.*

The estimate $\hat{\boldsymbol{\varpi}}^*$ can be obtained using a variable selection method that is consistent or possesses the screening property, such as SCAD[29], MCP[30], and Lasso[31;32]. If the sample size is reasonably large, one can even employ a sure independence screening method[33] here. The estimates $\hat{\xi}_j$'s can be obtained using a graphical modeling method, such as ψ-learning[23], nodewise regression[34], and graphical Lasso[35], all of which satisfy the required neighborhood screening property. However, compared to the latter two methods, ψ-learning provides a more explicit way for controlling the false discovery rate of the procedure. The MNR method is summarized in Algorithm 1.

In addition to the linear regression (5), the MNR method can be easily extended to other models. For instance, Liang et al[21] have extended the method to logistic regression and Cox regression, assuming that the explanatory variables follow a multivariate Gaussian distribution. More recently, Sun and Liang[36] relaxed the normality assumption for explanatory variables, enabling the method to be applied to mixed data.

3 Effective Connectivity Modeling via Time-Varying Dynamic Bayesian Networks

Task-based fMRI data modeling is an important problem in fMRI studies. During the past two decades, it has been addressed by quite a few groups of authors.[4;8;15;37;38] A popular way is to decompose the fMRI signal into two parts, activation, and effective connectivity. Mathematically, they expressed the model as

$$\boldsymbol{Y}_t^{(i)} = \boldsymbol{\mu}^{(i)} + \sum_{k=1}^{K} \boldsymbol{W}_k(t) \circ \boldsymbol{\gamma}_k^{(i)} + \boldsymbol{X}_t^{(i)}, \quad i = 1, 2, \ldots, n_t, \tag{9}$$

Algorithm 1 Markov Neighborhood Regression

Input: The random design matrix $\boldsymbol{X}$ and the response vector Y for a high-dimensional regression.

1. (**Variable selection**) Conduct variable selection for the model (1) to get a consistent estimate of $\boldsymbol{\varpi}_*$. Denote the estimate by $\hat{\boldsymbol{\varpi}}_*$.

2. (**Markov blanket estimation**) Construct a graphical model for $\boldsymbol{X}$ and obtain a consistent estimate of the Markov blanket for each variable. Denote the estimates by $\hat{\xi}_j$ for $j = 1, 2, \ldots, p$.

3. (**Subset regression**) For each variable X_j, $j = 1, \ldots, p$, let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{\boldsymbol{\varpi}}_*$ and run an ordinary least square (OLS) regression with the features given by $\boldsymbol{X}_{D_j}$, i.e.,

$$Y = \beta_0 + \boldsymbol{X}_{D_j} \boldsymbol{\beta}_{D_j} + \boldsymbol{\epsilon},$$

   where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$ and I_n is an $n \times n$-identity matrix. Conduct inference for β_j, including the estimate, confidence interval, and p-value, based on the output of the subset regression.

Output: The p-values for each variable X_j for $j = 1, 2, \cdots, p$.

where $\boldsymbol{Y}_t^{(i)} \in \mathbb{R}^p$ represents the BOLD signal collected from subject i at time t, p denotes the number of ROIs, $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^p$ represents the baseline mean value, $\boldsymbol{W}_k(t) \in \mathbb{R}^p$ represents a design vector for the k-th stimulus and is common for all subjects, $\boldsymbol{\gamma}_k^{(i)} = (\gamma_{1k}^{(i)}, \cdots, \gamma_{pk}^{(i)})^T$ is a p-vector representing the stimulus-specific regression coefficients for subject i, $\circ$ denotes Hadamard product, and $\boldsymbol{X}_t^{(i)} \in \mathbb{R}^p$ represents Gaussian random error from which effective connectivity can be inferred. As a generic notation for all subjects, we express $\boldsymbol{X}_t$ as $\boldsymbol{X}_t = (X_{t,1}, X_{t,2}, \ldots, \ldots, X_{t,p})^T$, where $X_{t,v}$ represents the Gaussian random error variable for the v-th ROI at time t. For convenience, we also refer to $X_{t,v}$ as an ROI.

In the model (9), the baseline mean value $\boldsymbol{\mu}^{(i)}$ describes the baseline signal without any stimulus for subject i. As in other works[4;8;15], we assume that the BOLD signal is characterized by a hemodynamic delay function (HRF), which explains the lapse of time between stimulus initiation and vascular response.[14] Therefore, we can model each element of $\boldsymbol{W}_k(t) = (W_{1k}(t), W_{2k}(t), \ldots, W_{pk}(t))^T$ as the convolution of the stimulus pattern with a HRF $h_r(t)$, i.e.,

$$W_{rk}(t) = \int_0^t w_k(\tau) h_r(t-\tau) d\tau, \quad t = 1, 2, \cdots, T, \quad r = 1, 2, \cdots, p, \tag{10}$$

where $w_k(\tau)$ represents the external time-dependent stimulus function for the k-th stimulus. In this paper, the canonical HRF is used, which is a common choice for modeling motor-based, visual-based, and emotion-based fMRI data.[14;15;38;39] The random error $\boldsymbol{X}_t^{(i)}$ contains the effective connectivity information and can be spatially and temporally dependent.

We propose to model $\boldsymbol{X}_t^{(i)}$ by a time-varying dynamic Bayesian network, given by

$$\boldsymbol{X}_t^{(i)} = \sum_{l=1}^{L} \boldsymbol{A}_{t,l} \boldsymbol{X}_{t-l}^{(i)} + \boldsymbol{\epsilon}_i(t), \tag{11}$$

where $\boldsymbol{\epsilon}_i(t) \sim N_p(0, \Sigma)$, $\Sigma = \text{diag}\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\}$ is a $p \times p$-diagonal matrix, l represents the lagging order, and $\boldsymbol{A}_{t,l} \in \mathbb{R}^{p \times p}$ is a time-varying transition matrix for time t. It is important to note that the autoregressive structure of (11) implies that different components of $\boldsymbol{X}_t^{(i)}$ can be generally dependent, although they are mutually independent conditioned on $\{\boldsymbol{X}_{t-l}^{(i)} : l = 1, 2, \ldots, L\}$. Through the estimation of the time-varying transition matrices $\boldsymbol{A}_{t,l}$'s, we will be able to infer the interactions among the ROIs and the information flow among brain regions, helping to uncover the working mechanism of the human brain. In

our analysis, we will first estimate γ_{ik}'s in equation (9) using the method by Friston et al,[40] thereby regressing out the strength of the activation term, and then learn the time-varying dynamic Bayesian network in equation (11) based on the residual term. Friston et al's method can be used for solving general linear models that allow for autocorrelations in the error terms, and it works well for fMRI data.

A Bayesian network is a directed acyclic graph (DAG) that can be used to infer the directed structure of a dynamic system consisting of a large set of random variables. However, during the past two decades, most research on Bayesian networks has focused on static networks, which involves inferring the network structure using data collected at a single time point[41–43]. In practice, there is often a need to infer the directed structure for a dynamic system. For this purpose, the time-varying dynamic Bayesian network becomes a powerful tool. It allows for the inference of the evolving structure of the system over time, accommodating the changing relationships between variables.

For the time-varying dynamic Bayesian network, or more generally, a dynamic DAG, we assume that it satisfies the directed Markov property and faithfulness[44]. The dynamic DAG, as defined by the model (11), is said to satisfy the directed Markov property if and only if each ROI $X_{t,v}$ in $\boldsymbol{X}_t$ and its older ancestors in $\{\boldsymbol{X}_k : k = t-L-1,\ldots,1\}$, denoted by $\boldsymbol{X}_{t,an(v)\setminus pa(v)}$, are conditionally independent given its parents in $\{\boldsymbol{X}_k : k = t-1,\ldots,t-L\}$, denoted by $\boldsymbol{X}_{t,pa(v)}$, i.e.,

$$X_{t,v} \perp\!\!\!\perp \boldsymbol{X}_{t,an(v)\setminus pa(v)} | \boldsymbol{X}_{t,pa(v)}.$$

Thus, the dynamic DAG prescribes a set of conditional independence relations on the ROIs comprising the graph. Faithfulness ensures that we can read off all conditional independence relations from the graphical concept of separation. In addition to the dynamic DAG, we have also assumed the Markov property and faithfulness properties for the Gaussian graphical models for the ROIs at each time point t. See Liang et al.[23] for further discussions on this issue. In summary, the Markov property and faithfulness assumptions warrant a systematic statistical framework for the proposed dynamic DAG to infer the effective connectivity of ROIs using fMRI data.

4 Learning Time-Varying Dynamic Bayesian Networks

Consider a sequence of datasets: $\{\boldsymbol{X}_t^{(i)} : i = 1, 2, \ldots, n_t\}$ for $t = 1, 2, \ldots, T$, where $\{\boldsymbol{X}_t^{(i)} : t = 1, 2, \ldots, T\}$ forms a time series collected from subject 'i'. However, for each subject, it is not required to be observed at all time points. As a result, the sample size 'n_t' can be different at different time points. For the time being, we assume that each observation $\boldsymbol{X}_t^{(i)}$ is drawn from a Gaussian distribution. The extension of the proposed methods to other distributions will be discussed at the end of the paper.

To learn a time-varying dynamic Bayesian network from the datasets, we propose a two-stage method. The first stage is to jointly estimate multiple Gaussian graphical models across all T time points using an accelerated hybrid Bayesian integrative analysis method[22]; and the second stage is to infer the time-varying dynamic Bayesian network using Markov neighborhood regression (MNR)[21], where the Markov neighborhood of each variable is formed based on the multiple Gaussian graphical models obtained in the first stage. For simplicity, we coined the proposed method as 'joint Gaussian graphical model plus Markov neighborhood regression' or JGGM+MNR in short. Figure 1 depicts the pipeline of the proposed method. The details are provided in the following subsections.

Notation: In the rest of the paper, we will use $\boldsymbol{X} = \{X_1, X_2, \cdots, X_p\}$ to represent p variables or nodes of a Bayesian network. Conceptually, $\boldsymbol{X}_t^{(i)}$ can be regarded as a realization of $\boldsymbol{X}$.

4.1 Joint Estimation of Multiple Graphical Models

We have developed an accelerated hybrid Bayesian integrative analysis method, which is an extension of Jia et al,[22] for jointly estimating multiple Gaussian graphical models at a large number of time points. The proposed method consists of three steps: edgewise score evaluation, Bayesian data integration, and joint edge detection. In the first step, we evaluate a conditional independence test score for each potential edge in the multiple graphical models. Then, in the second step, we refine the conditional independence test scores within a Bayesian framework by integrating data information from different time points. Finally, in the third step, we jointly test the significance of the refined scores using a multiple-hypothesis testing procedure. The first and third steps follow Jia et al[22]; while in the second step, we utilize

a stochastic EM algorithm[45;46] for Bayesian edge clustering. This extension enables us to handle problems with a large number of time points. We will provide a detailed description for each of the three steps in what follows.

4.1.1 Edgewise Score Evaluation

In this step, a conditional independence test score is evaluated for each potential edge in the multiple graphical models at each time point. Since the assessments are performed separately for each time point, we will describe the procedure for a single time point for the sake of simplicity, with the time point index suppressed.

Consider a graphical model with p nodes given by $\mathbf{X} = \{X_1, X_2, \cdots, X_p\}$. Suppose that the graphical models satisfy the Markov and faithful properties. Let $\boldsymbol{S}_j$ denote the Markov blanket of X_j. As implied by the total conditioning property[47] of Markov blankets, the two events $X_j \perp X_i | \boldsymbol{X} \backslash \{X_i, X_j\}$ and $X_j \perp X_i | \boldsymbol{S}_j \backslash \{X_i, X_j\}$ are equivalent in the sense

$$\delta_{ij}^V = 1 \Longleftrightarrow \delta_{ij} = 1, \tag{12}$$

where δ_{ij}^V and δ_{ij} denote, respectively, the indicator functions of $X_j \perp X_i | \boldsymbol{X} \backslash \{X_i, X_j\}$ and $X_j \perp X_i | \boldsymbol{S}_j \backslash \{X_i, X_j\}$. In a graphical model, for any random variable X_j, its Markov blanket is any subset $\boldsymbol{S} \subset \mathbf{X} \setminus \{X_j\}$ conditioned on which other variables are independent of X_j. If $\boldsymbol{S}_j$ is replaced by a Markov blanket of X_i, denoted as $\boldsymbol{S}_i$, (12) also holds. If X_i and X_j are independent conditioned on $\boldsymbol{S}_j \backslash \{X_i, X_j\}$, then there is no link between X_i and X_j. Otherwise, X_i and X_j are linked in the graph.

Let $\tilde{\boldsymbol{S}}_j (\supset \boldsymbol{S}_j)$ represent a super-Markov blanket of X_j. In general, the super-Markov blanket can be obtained by a simple and fast marginal correlation screening method, and its size can be limited to $\mathcal{O}(n/\log(n))$.[33] Let $\tilde{\delta}_{ij}$ denote the indicator function of the event $X_j \perp X_i | \tilde{\boldsymbol{S}}_j \backslash \{X_i, X_j\}$. Under the faithfulness assumption, Xu et al[48] showed that the two events $X_j \perp X_i | \boldsymbol{S}_j \backslash \{X_i, X_j\}$ and $X_j \perp X_i | \tilde{\boldsymbol{S}}_j \backslash \{X_i, X_j\}$ are equivalent in the sense

$$\delta_{ij} = 1 \iff \tilde{\delta}_{ij} = 1.$$

Also, they suggested using the Sure Independence Screening (SIS) approach[49] to determine the super-Markov blanket for each node. That is, for each variable, SIS is conducted with

respect to the remaining variables and the selected variables from its super-Markov blanket. This leads to Algorithm 2 for edgewise score evaluation. Since both the marginal correlation screening and the conditional independence test only need to be performed once for each ordered pair of nodes. Therefore, the computational complexity of Algorithm 2 is $\mathcal{O}(p^2)$, which remains independent of the underlying structure of the graphical model.

Other than the SIS approach, the nodewise regression approach[34] can also be used for the construction of super-Markov blankets due to the screening properties of the Lasso method. In applications, Algorithm 2 can be implemented using the R package SIS.[51]

4.1.2 Bayesian Data Integration

As mentioned previously, the proposed method works under the scenario that multiple observations of $\boldsymbol{X}_t$ are available at each time point. That is, the observations form a three-dimensional set $\{\mathcal{X}_1, \mathcal{X}_2, \cdots \mathcal{X}_T\}$, where T denotes the total number of different time points, $\mathcal{X}_t = \{\boldsymbol{X}_t^{(1)}, \boldsymbol{X}_t^{(2)}, \ldots, \boldsymbol{X}_t^{(n_t)}\}$ denotes n_t observations of $\boldsymbol{X}_t$, and each observation $\boldsymbol{X}_t^{(i)} \in \mathbb{R}^p$ (for $i = 1, 2, \ldots, n_t$) is a p-dimensional vector.

Suppose z-scores have been calculated for each possible edge at each time point by using Algorithm 2. Then we have the scores stored in an $N \times T$-matrix $\boldsymbol{Z} = (z_{ij})$, where $N = p(p-1)/2$ denotes the number of z-scores calculated at each time point. In the following, we describe how the scores are fine-tuned by integrating information from the data collected at different time points under the Bayesian framework.

For each edge l, we denote the vector $\boldsymbol{e}_l = (e_l^{(1)}, e_l^{(2)}, \ldots, e_l^{(T)})^T$ as the corresponding status (existence or absence) at T time points. Specifically, $e_l^{(t)} = 1$ indicates the existence of the edge at time t, and 0 otherwise. To enhance the similarity of the edge across different time points, we impose a temporal prior distribution on $\boldsymbol{e}_l$:

$$\mathbb{P}(\boldsymbol{e}_l|q) = \frac{1}{2} q^{\sum_{t=1}^{T-1} 1-c_t} (1-q)^{\sum_{t=1}^{T-1} c_t},$$

where we set the prior probability $\mathbb{P}(e_l^{(1)} = 1) = \mathbb{P}(e_l^{(1)} = 0) = \frac{1}{2}$; $c_t = |e_l^{(t+1)} - e_l^{(t)}|$, for $t = 1, 2, \cdots, T-1$, indicates the change of the status of the edge l from the time point t to the time point $t+1$, and q is the prior probability of an edge maintaining its status unchanged from one time point to the next. In this paper, we assume that q follows a

Beta(a_1, b_1) distribution, where $a_1 = 10$ and $b_1 = 1$ are the default values specified for the prior hyperparameters. This setting results in a prior mean of 0.909 for q, enhancing the structural stationarity of the Bayesian networks across different time points.

Following Jia et al,[52] we assume that the z-scores $z_l^{(t)}$'s are mutually independent conditioned on $\boldsymbol{e}_l$, and each follows a two-component mixture Gaussian distribution,

$$\mathbb{P}(z_l^{(t)}|e_l^{(t)}) = \begin{cases} \mathrm{N}(\mu_{l0}, \sigma_{l0}^2) & \text{if } e_l^{(t)} = 0, \\ \mathrm{N}(\mu_{l1}, \sigma_{l1}^2) & \text{if } e_l^{(t)} = 1, \end{cases}$$

for $l = 1, 2, \cdots N$ and $t = 1, 2, \cdots T$. That is, we will cluster the z-scores into two clusters, corresponding to the cases where the edge exists and where the edge does not exist, respectively. Let $\boldsymbol{z}_l = \{z_l^{(1)}, z_l^{(2)}, \cdots z_l^{(T)}\}$. The conditional likelihood function of $\boldsymbol{z}_l$ is given by

$$\mathbb{P}(\boldsymbol{z}_l|\boldsymbol{e}_l) = \prod_{\{t:e_l^{(t)}=0\}} \phi(z_l^{(t)}|\mu_{l0}, \sigma_{l0}^2) \prod_{\{t:e_l^{(t)}=1\}} \phi(z_l^{(t)}|\mu_{l1}, \sigma_{l1}^2), \tag{13}$$

where $\phi(\cdot|\mu, \sigma^2)$ is the density function of the Gaussian distribution with mean μ and variance σ^2. Taking a product of the likelihood function (13) over all edges $l = 1, 2, \cdots N$, we will get the joint likelihood function for all z-scores. In some cases, a three-component mixture Gaussian distribution might be required to model the z-scores. For example, when an edge is present in multiple graphs, but its partial correlation coefficients have different signs across those graphs. As elaborated in Jia et al,[22] this extension can be easily implemented.

Jia et al[22] considered the case that T is small. In this case, by integrating other parameters, the marginal posterior probability of each of 2^T possible configurations of $\boldsymbol{e}_l$ can be evaluated exactly. For the case that T is large, they suggest an MCMC approach, but which is too slow. Inspired by the imputation-regularized optimization (IRO) algorithm,[53] we propose a new Bayesian clustering method for simulating samples of $\boldsymbol{e}_l$, with the detail given by Algorithm 3.

Let $\theta_{l,k} = (\hat{\mu}_{l0}^{(k)}, \hat{\sigma}_{l0}^{(k)}, \hat{\mu}_{l1}^{(k)}, \hat{\sigma}_{l1}^{(k)})$ denote the parameter estimate obtained for the edge l at iteration k of Algorithm 3. Similarly, we let $\boldsymbol{e}_{l,k}$ denote the sample of $\boldsymbol{e}_l$ imputed at iteration k of the algorithm. For each edge l, Algorithm 3 leads to two interleaved Markov chains

$$\theta_{l,0} \to \boldsymbol{e}_{l,0} \to \theta_{l,1} \to \boldsymbol{e}_{l,1} \to \cdots \to \theta_{l,k} \to \boldsymbol{e}_{l,k} \to \cdots,$$

which can converge very fast, usually within tens of iterations. According to the convergence theory established by Liang et al[53], under mild conditions, $\theta_{l,k}$ will converge in probability to $\theta_{l,*}$, the maximum *a posteriori* (MAP) estimate of $(\mu_{l0}, \sigma_{l0}, \mu_{l1}, \sigma_{l1})$; and $\boldsymbol{e}_{l,k}$ will converge weakly to the posterior distribution $\pi(\boldsymbol{e}_l|\theta_{l,*}, \boldsymbol{z}_l)$. In this paper, for simplicity, we assume that the parameters $(\mu_{l0}, \sigma_{l0}, \mu_{l1}, \sigma_{l1})$ are subject to an improper prior $\pi(\mu_{l0}, \sigma_{l0}, \mu_{l1}, \sigma_{l1}) \propto 1$, resulting in the maximum likelihood estimates for these parameters within each cluster. For cases that T is small, we recommend Jeffreys' prior $\pi(\mu_{l0}, \sigma_{l0}, \mu_{l1}, \sigma_{l1}) \propto 1/(\sigma_{l0}^2\sigma_{l1}^2)$ or the inverse Gamma prior as used in Jia et al.[22]

Let $\boldsymbol{e}_{l,1}, \boldsymbol{e}_{l,2}, \ldots, \boldsymbol{e}_{l,m}$ denote m samples simulated using Algorithm 3 for edge l. For each sample $\boldsymbol{e}_{l,j}$, we calculate the integrated z-scores using Stouffer's meta-analysis method.[54;55] That is, for $t = 1, 2, \ldots, T$, we define the integrated z-score as

$$\bar{z}_{l,j}^{(t)} = \begin{cases} \sum_{\{i:e_{l,j}^{(i)}=0\}} w_i z_{l,j}^{(i)} / \sqrt{\sum_{\{i:e_{l,j}^{(i)}=0\}} w_i^2} & \text{if } e_{l,j}^{(t)} = 0, \\ \sum_{\{i:e_{l,j}^{(i)}=1\}} w_i z_{l,j}^{(i)} / \sqrt{\sum_{\{i:e_{l,j}^{(i)}=1\}} w_i^2} & \text{if } e_{l,j}^{(t)} = 1, \end{cases}$$

where the weights w_i's can be assigned to account for the time-varying property of the dynamic system. For example, they can be assigned according to the size or quality of the data collected at different time points. In this paper, we set $w_i = 1$ for $i = 1, 2, \cdots T$ in our numerical studies. Then the Bayesian integrated z-scores, also called Bayesian Stouffer integrated z-scores, are given by

$$\hat{z}_l^{(t)} = \sum_{j=1}^{m} \bar{z}_{l,j}^{(t)}/m, \quad l = 1, 2 \cdots, N \text{ and } t = 1, 2, \cdots, T.$$

Alternative to Stouffer's meta-analysis method, Fisher's method and Pearson's method[56] can also be applied here with minor modifications.

4.1.3 Joint Edge Detection

In this step, we conduct multiple hypothesis tests to identify the Bayesian integrated z-scores that are significantly larger than the others and set the edges accordingly. More precisely, we use $\hat{z}_l^{(t)}$ as the test statistic for the existence of edge $l \in \{1, 2, \ldots, N\}$ at time $t \in \{1, 2, \ldots, T\}$ in the multiple hypothesis test. In this paper, we adopted the empirical Bayesian method,[57] which allows general dependency between the test statistics. More

related research on empirical Bayesian methods in the context of multiple hypothesis testing includes references[58–60].

4.2 Time-Varying Dynamic Bayesian Network Learning

Utilizing the methods of joint estimation of multiple graphical models and the MNR method, the latter introduced in Section 2 and outlined in Algorithm 1, we can construct the time-varying dynamic Bayesian network as described in Algorithm 4.

Algorithm 4 contains two important free parameters, namely, the significance levels used in two multiple-hypothesis tests. One test is for joint edge detection in the step of 'joint estimation of multiple graphical models', and the other test is for the step of 'joint link detection'. We denote the significance level of the former test by α_1 and that of the latter by α_2. For both tests, we adopt the empirical Bayesian method,[57] where the significance of the test is measured in Storey's q-value.[61] To ensure the minimum Markov neighborhood ξ_j is covered by D_j for each node j and at each time point t, we suggest a large value of α_1, e.g., 0.1 or 0.2, to be used. In this paper, the default values $\alpha_1 = 0.2$ and $\alpha_2 = 0.05$ are used unless stated otherwise. We note that α_1 and α_2 play the same role as regularization parameters in a regularization method. Specifically, they function as the regularization parameter in a JGGM method: α_1 controls the size of the estimated Markov neighborhoods, while α_2 controls the sparsity of the resulting multiple Bayesian networks. It is evident that a large value of α_1 reduces the risk of missing important features in the estimated Markov neighborhood, but induces higher uncertainty in the subsequent MNR estimation.

Tips for Tuning the values of α_1 and α_2. As α_1 governs the size of the Markov neighborhoods used in the MNRs, we recommend selecting its value to ensure that each MNR encompasses some features with big p-values, which implies the Markov neighborhood has been sufficiently large to include some false features. Sensitivity studies have been conducted with respect to α_1 for a simulated example as well as the real data set, see the supplement Table S8 and Figures S8 for the detail. The choice of α_2 can be simply determined based on the desired density of edges in the resulting Bayesian networks.

In addition to α_1 and α_2, the algorithm contains a few other parameters, including the

parameters used in sure independence screening for super-Markov blankets construction, the number of iteration steps used in Bayesian clustering, and the algorithm and parameters used in high-dimensional variable selection in the MNR step. In this paper, we adopted the SIS method[33] with ℓ_1 penalty to estimate the super-Markov blanket for each node, which was implemented using the R-package *SIS*[51] under their default settings. For Bayesian edge clustering, we set the default number of iterations to $\mathcal{K} = 200$, where the samples produced in the last 20 iterations were used for inference.

Algorithm 4 is highly attractive in computation. In the step of 'joint estimation of multiple graphical models', the edgewise scores can be computed in parallel for all edges, and Bayesian edge clustering and data integration can be performed in parallel for all edges as well. Similarly, the step of Markov neighborhood regression can be executed in parallel for all nodes at each time point $t = 2, 3, \ldots, T$. For joint edge detection and joint link detection, the empirical Bayesian method[57] utilizes the stochastic gradient descent (SGD) method to estimate the mixture distribution of the z-scores, which is subsequently used for false discovery rate (FDR) evaluation. Therefore, these two steps can be accelerated using the mini-batch strategy.

Remark 4.1. *In the case that the Gaussian random noise in the model (11) is spatially dependent, i.e., Σ is non-diagonal, the MNR method for the high-dimensional regression (14) of Algorithm 4 only needs to be slightly modified by further expanding the super Markov blanket of each variable $X_{t-1,k}$ (for any $k \in \{1, 2, \ldots, p\}$) to include the neighboring variables of $X_{t,j}$ in the graph $\mathcal{G}_t$. This construction of the MNR can be justified by the structural equation representation of a Gaussian graphical model, see e.g. Bühlmann and van de Geer (p.436).*[32] *Note that in this case the edges learned from the graph $\mathcal{G}_t$ are no longer directed, and the resulting network cannot be interpreted as a Bayesian network anymore. Refer to Section S2.3 of the supplement for a numerical study for this extension of the method.*

Remark 4.2. *In this paper, we have assumed that the sample size n_t can be different at different time points. However, the MNR step involves a subject-matching issue in forming the high-dimensional regression. Therefore, for the MNR step, only a subset of the samples taken from the common subjects at time $t-1$ and t can be used. Allowing different sample*

sizes at different time points enables us to use more data for learning the graphical models $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T$.

Remark 4.3. *The consistency of Algorithm 4 for dynamic Bayesian network learning can be justified based on the consistency of the MNR method*[21] *for identifying the directed structure around the response variable. Refer to Section S1 of the supplement for the details.*

5 Synthetic Examples

This section demonstrates the effectiveness of the proposed method using two synthetic examples. The first example showcases the performance of the new method compared to the Bayesian nodewise selection (BANS) method[62] on a small-scale dataset. The BANS method was originally designed for learning multilayered Gaussian graphical models, with directed edges between layers and undirected edges within the same layer. Consequently, it can be applied to learn time-varying dynamic Bayesian networks within the context of this paper. BANS employs Bayesian variable selection priors for each of the regressions to jointly select both undirected and directed edges that point to a node. BANS is widely recognized as a state-of-the-art method for learning time-varying dynamic Bayesian networks. To demonstrate the proposed method's ability to model whole-brain networks, we test it on high-dimensional and large-scale datasets in the second example. The sensitivity analysis for the proposed method is provided in the supplement. For each synthetic example, the proposed method was tested on 10 independent datasets and run in parallel with 20 threads on a personal computer equipped with an i9-10900K CPU@3.6GHz and 128GB of memory.

5.1 A Comparison with the BANS method

To mimic the real task-based fMRI data, we generated data from equations (9) to (11) with a lag order of $L = 1$, where γ_{irk} for $r = 1, 2, \ldots, p$ is drawn from the uniform distribution $\mathcal{U}(0, 5)$, $\boldsymbol{\mu}^{(i)} = 1$ for each subject, $\boldsymbol{\epsilon}_i(t) \sim N_p(0, \Sigma)$ with $\Sigma = \text{diag}\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\}$ and each σ_j^2 is drawn from the uniform distribution $\mathcal{U}(0.9, 1.2)$. See Section S2.1 of the supplement for the detailed setting of data generation.

During the data analysis stage, we first eliminated the strength of activation for the nodes through regression and then proceeded to learn the dynamic Bayesian network with the residuals (11). To evaluate the performance of the methods, we used the precision-recall curve. The results are summarized in Table 1 and Figure 2. We considered $n = 400$ and different numbers of nodes ($p = 20, 30, 50$). Compared to the BANS method, the proposed method is not only superior in the identification of dynamic network structure but also computationally more efficient. The performance of the proposed method was further assessed in the supplement under other scenarios, including weak signals, spatial correlations, and varying connection strengths for different subjects. In all of these cases, the proposed method consistently outperforms the BANS method.

Time Complexity We approximated the time complexity of the proposed method by a linear regression model with respect to the sample size n and the node size p. Under the settings $n = 200, 300, 400$ and $p = 16, 20, 24, 30, 40$, and $T = 60$, we ran the proposed method and BANS to acquire their respective computational time. Figure 3 presents the curves for the computational time when the number of variables p varies, where different colors represent different sample sizes. When p is fixed and n increases, the computational time of the new method does not vary much, while that of BANS increases significantly. Please refer to the supplement for the details of the experiments. This study shows that the BANS method is only capable of learning small dynamic Bayesian networks when T is large.

5.2 High-Dimensional and Large-Scale Data Case

In this subsection, we conducted simulation studies under the large p and large n scenarios to demonstrate the potential of the proposed method for learning whole brain networks. The data generation procedures and simulation settings are the same as described in Section 5.1. Two cases were considered here: (i) $n = 800$, $p = 300$, and $T = 60$, (ii) $n = 400$, $p = 500$ and $T = 60$. Case (i) represents a large dataset, whose scale is close to a real fMRI dataset, while case (ii) represents a small-n-large-p dataset. In case (ii), to illustrate the robustness of the proposed method, we simulated the transition matrices with varying connection strengths for different subjects (see Section S2.4 of the supplement). For both cases, we simulated

data under the strong signal setting.

For both cases, we compared the proposed method with some existing regularization methods, including Lasso,[31] Elastic Net,[63] and MCP,[30] which fit a high-dimensional linear regression separately for each ROI from $t-1$ to t. The numerical results are summarized in Figure S1 (in the supplement), which suggests that the proposed JGGM+MNR method outperforms regularization methods significantly in both cases. The existing literature consistently shows that the MNR exhibits superior performance in variable selection compared to regularization methods, see references[21;36;50] for the detail. Furthermore, in stage one, the proposed method provides an accurate estimator $\hat{\boldsymbol{S}}_{j,t}$ for the Markov neighborhood of each node j at each time point t.

6 An fMRI Study of Emotion Processing

This section presents the results of the proposed method in an fMRI study of emotion processing. Unfortunately, the BANS method does not apply to this study due to its high time complexity. As a means of partially validating our results, we carried out sensitivity analyses, which are detailed in the supplement.

6.1 Data Acquisition and Pre-processing

The data set we used is publicly available from the S1200 Data Release of the Human Connectome Project (HCP).[64] We specifically selected the subjects who had complete data for all four 3T MRI modalities in the HCP protocol: structural images (T1w and T2w), resting-state fMRI (rsfMRI), task fMRI (tfMRI), and high angular resolution diffusion imaging (dMRI). A total of 867 subjects aged between 22 and 35 years were included in our analyses, comprising 409 males and 458 females. The fMRIs were acquired using a whole-brain multi-band gradient-echo (GE) echoplanar (EPI) sequence with the following parameters: TR/TE = 720/33.1 ms, flip angle = 52°, FOV = 208 × 180 mm, matrix = 104 × 90 (RO × PE), multi-band factor = 8, echo spacing = 0.58 ms, and slice thickness = 2 mm. The resulting normal voxel size was 2.0 × 2.0 × 2.0 mm. Additional information about the S1200 Data Release of HCP can be found in the reference manual at `https://www.humanconnectome.org`.

Specifically, we focused on the emotion processing task fMRI with left-to-right encoding. The task was adapted from the study conducted by Hariri et al. (2006)[24] As illustrated by Figure 4, the participants underwent six blocks of trials, consisting of either face or shape matching. Among these blocks, three were face blocks, and three were shape blocks. The faces displayed either angry or fearful expressions. Each block lasted for 21 seconds, including a cue phase with six repetitions of the same task, followed by a 2-second stimulus presentation and a one-second inter-trial interval (ITI). The total duration of the task was 2 minutes and 16 seconds, with 176 frames acquired for each subject.

The fMRI we used from the HCP has undergone a minimal pre-processing pipeline, which encompassed gradient distortion correction, head motion correction, image distortion correction, spatial normalization to the standard Montreal Neurological Institute (MNI), and intensity normalization.[65] Additionally, we implemented standard pre-processing steps to reduce biophysical and other noise in the minimally processed fMRI data. These procedures involved eliminating the linear components associated with the 12 motion parameters (including the original motion parameters and their first-order derivatives), removing linear trend, and applying band-pass filtering ($0.01 - 0.25$ Hz).[66] The band-pass filtering with a range of $0.01 - 0.25$ Hz was applied to achieve two objectives in this data: 1) eliminating the scanner drift and linear trends by high-pass filtering, and 2) attenuating respiratory noise and cardiac signal through low-pass filtering.[67]

For a better understanding of the behavior of different brain regions, we extracted 268 regions of interest (ROIs) based on the functional atlas provided by Finn et al. (2015).[66] We averaged the values of all voxels that belonged to each ROI. Furthermore, we divided the 268 ROIs into 8 functional networks (FN) modules (see Figure S7 in the supplement) including Medial Frontal (Med F), Frontoparietal (FP), Default Mode (DMN), Subcortical-cerebellum (Sub-Cer), Motor (Mt), Visual I (Vis I), Visual II (Vis II) and Visual Association (Vis Assn) as in Cai et al.[67]

6.2 Parameter Settings

In the joint estimation of multiple graphical models, we set $\alpha_1 = 0.2$. Then, we tried $\alpha_2 = 0.05$ and 0.1 to select important variables for constructing the time-varying dynamic

Bayesian network. We considered the lag orders $L = 1$ and $L = 2$. We summarized the results with $\alpha_2 = 0.1$ and $L = 1$ in the main paper and left other results to the supplement.

6.3 Results

We have acquired 175 dynamic DAGs. Within each block, the task duration was $18s$, resulting in 25 DAGs. Therefore, we obtained 75 DAGs to describe the directed functional connectivity (FC) in emotion processing (case), and 75 estimations for the shape processing (control). To better highlight the differences between the emotion and shape tasks, we summarized the networks from four aspects, namely characteristic edges, functional network modules, significant ROIs (hubs), and network differences.

6.3.1 Task-related Networks under Emotion Processing and Shape Processing

In Figure 5, we present the chord plots of the mean of the dynamic DAGs over time, arranged by functional network modules, and the characteristic edges for each task. A characteristic edge is defined as an edge whose total number of appearances is at least two standard deviations higher than the mean number (which is 9 for both tasks). The chord plots depicting the functional network modules for the two tasks exhibit overall similarities, with variations in inter-modular connectivity and appearance at different locations. Regarding the characteristic edges, we identified 14 edges for emotion processing and 10 edges for shape processing. Among them, 6 edges were found in both tasks, which are the connectivity between the left and right hemispheres in the motor strip and limbic areas, as well as connectivity involving the parietal lobe. Notably, the cerebellum demonstrated more activity during emotion processing compared to shape processing.

6.3.2 Emotion-Related Intra- and Inter-Modular Connectivity

We examined intra-modular and inter-modular connectivity for the emotion and shape tasks, as well as the differences between the two conditions (see Figure 6). In the heat maps, the rows indicate the beginning of the arrows, and the columns indicate the end of the arrows. The significant differences in intra-modular and inter-modular connectivity depicted

in Figure 6 were selected based on the visualized heat map and the statistical modeling analysis.

By considering the difference in mean edge degrees between the emotion and shape tasks and the sensitivity result of the Poisson HRF (available in the supplement), we identified some differences in the intra- and inter-modular connectivities under the two tasks.

In the emotion task, compared to those in the shape task, the intra-modular connectivity in the Subcortical-cerebellum networks is less active, while the inter-modular connectivity among the Subcortical-cerebellum networks, motor networks, visual II, and visual association networks is more active.

The results (see Figure 5) confirmed the central role of the Subcortical-cerebellum area in the coordination of Motor behavior and cognitive processing.[68] The study found that in the Subcortical-cerebellum module, there is a significant difference between the emotion and shape tasks (see Figure 6). Subcortical structures refer to a group of diverse neural formations located deep within the brain, including the diencephalon, pituitary gland, limbic system, and basal ganglia. They are involved in complex activities such as memory, emotion, pleasure, and hormone production. The limbic system regulates autonomic and endocrine functions in response to emotional stimuli.[69] The results also provide evidence of the cerebellum's contribution to complex network processing, emotional facial expression, and enhanced emotional recognition of facial anger and sadness.[70] The cerebellum's connection to the emotional processing task has been well-studied in the literature.[71–73] Our results further showed that during emotion processing, there were more active connections in the pathways of Sub-Cer $\leftrightarrow$ Motor, Sub-Cer $\rightarrow$ Vis II, and Sub-Cer $\rightarrow$ Vis Assn than during shape processing.

6.3.3 Identification of Emotion-related ROIs

To gain more insights into the emotion circuits, an analysis was conducted on the degree distribution of the nodes, leading to the identification of a hub ROI that exhibited differential performance under the emotion and shape tasks. Row degrees were calculated to determine the total number of edges coming out of a node, and the significance of information transmission was assessed using the Wilcoxon signed-rank test. The test suggests that the row

degrees of ROI 32 vary significantly under different tasks. Figure S9 in the supplement displays this variation along with the time, which tends to have higher connectivity under the emotion task than under the shape task. Detailed information on the test can be found in the supplement. A literature review reveals that the region BA6, to which ROI 32 belongs, may play an important role in facial emotion processing.[74;75]

6.3.4 Network Comparison

To examine the structural variation of the dynamic Bayesian networks learned by the proposed method, a multiple-hypothesis test was conducted on the edges of the networks using the method of ψ-learning.[23] Specifically, considering two dynamic Bayesian networks under comparison, we define the test statistic for each pair of possible edges as $z_{ij} = \frac{z_{ij}^{(1)} - z_{ij}^{(2)}}{\sqrt{2}}$, where the subscript '$ij$' specifies a possible edge, and the superscript 'k' (for $k = 1, 2$) specifies a Bayesian network. The comparisons were done for the dynamic Bayesian networks learned from time t to $t+1$ and from time $t+1$ to $t+2$ within each of the emotion and shape tasks. At a significance level of 0.1, the proposed method identified 137 and 130 differential edges within the shape and emotion tasks, respectively, with $L = 1$; and it identified 22 and 37 differential edges within the shape and emotion tasks, respectively, with $L = 2$. Additionally, we examined the structural variation of the dynamic Bayesian network during the transitions between the emotion and shape tasks, identifying 24 and 9 differential edges with $L = 1$ and $L = 2$, respectively. This result suggests that the transition matrices $A_{t,l}$ are time-varying within each of the emotion and shape tasks, providing evidence to support our assumption of time-varying connectivity. Compared to the case of $L = 1$, fewer differences in connectivity were detected in the case of $L = 2$, which suggests a high-order Markovian nature of brain activity.

7 Discussion

In this paper, we have proposed a new method for learning time-varying dynamic Bayesian networks with multiple samples observed for a dynamic system at a large number of time points. We have established its consistency for dynamic Bayesian network construction. The

proposed method has been successfully applied to an fMRI study for emotion processing. We found that during the emotion task, the modular subcortical-cerebellum plays a key role, and significant differences are presented by comparing it to the shape task.

The proposed method provides a general framework for learning causal graphs from spatio-temporal data. It comprises two stages: JGGM, which extracts spatial-type structural information from the data; and MNR, which determines the structure of dynamic Bayesian networks by incorporating both the temporal information and the spatial information extracted by JGGM. We implemented the JGGM using the fast hybrid Bayesian integrative analysis (FHBIA) method[22]. As mentioned in the Introduction, several JGGM estimation methods are available in the literature, including fused Lasso[11;12] and Bayesian modeling via a Markov random field[13]. The former enforces temporal homogeneity through a fused Lasso penalty, but an issue with the method is its assumption of independence among observations at different temporal points, which is challenging to satisfy. The latter enforces temporal homogeneity using a Markov random field prior, but it involves repeated inversions of covariance matrices, rendering it impractical for large-size networks. The FHBIA method successfully addresses these issues. As described in the paper, it first employs the ψ-learning method[23], a frequentist method, to transform the original data into edge-wise z-scores. The z-score serves as an equivalent measure of the partial correlation coefficient, providing a concise summary of the graph structure information within the data under each condition. Subsequently, it employs a Bayesian method to model the z-scores for edge clustering and applies a meta-analysis technique to integrate data information across distinct conditions. Since FHBIA comprises both frequentist and Bayesian components, it is hybrid and so is the proposed one in the paper. Compared to the fully Bayesian method[13], the use of the z-score transformation allows us to avoid inversions of high-dimensional covariance matrices, significantly enhancing the computational efficiency of the proposed method. Compared to the fused Lasso method[11;12], the Bayesian edge clustering step enables the incorporation of prior information to enhance the temporal homogeneity flexibly, while accounting for the dependent nature of the data in temporal modeling. Furthermore, the MNR method provides a natural way to integrate spatial and temporal information in determining the structure of the dynamic Bayesian network.

The proposed method has a striking feature in uncertainty quantification. Specifically, it learns a z-score for each possible link, including both identified and non-identified links, in the dynamic Bayesian network. These z-scores allow us to assess the reliability of the learned network, a capability beyond the reach of regularization methods, such as applying the nodewise regression[34] to the problem.

The proposed method is also very attractive in computation, which has a time complexity of $O(p^2)$ under the assumptions $p \gg n$ and $p \gg T$. Furthermore, almost all steps of the method can be run in parallel, and the joint edge detection and joint link detection steps can be accelerated using the mini-batch strategy. An R-package has been developed for the proposed method, which can be run on a multi-core computer in parallel. The package will be made available to the public upon publication of the paper.

Finally, we note that a limitation of the current version of the proposed method is its requirement for the data to be multivariate Gaussian. However, extending the proposed method to mixed data is straightforward. In this case, the joint estimation of multiple graphical models can be done as described in Jia and Liang,[52], and Markov neighborhood regression for mixed data can be done as described in Sun and Liang.[36] Many of the assumptions used in the paper can also be relaxed. For example, the weight of each subject is not necessarily equal. In this case, although the effective connectivity is assumed to be equal across all subjects, the weight of each subject can be incorporated into the proposed method by replacing each Markov neighborhood regression with its weighted version (i.e., weighted linear regression). In addition, the stationarity of the time series is unnecessary due to the time-wise analysis nature of the proposed method. As a slight extension of the proposed model, task information can be used in both steps of graphical model construction and Bayesian network structure determination. For the former, task information can be used as an external variable, and ψ-scores can be computed by accounting for its effect as prescribed in Liang et al[23] For the latter, it can always be included as a predictor of each Markov neighborhood regression.

Acknowledgements

The fMRI Data were provided in part by the Human Connectome Project, WU-Minn Consortium (principal investigators, D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 US National Institutes of Health (NIH) institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Liang's research was supported in part by the NSF grants DMS-2015498 and DMS-2210819 and NIH grant R01-GM126089. The authors thank the editor, associate editor, and two referees for their insightful and constructive comments, which have led to significant improvement of this paper.

References

[1] Ogawa Seiji, Lee T M, Kay A R, Tank David W.. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America.* 1990;87 24:9868-72.

[2] Petersen Steven E, Sporns Olaf. Brain networks and cognitive architectures *Neuron.* 2015;88:207–219.

[3] Bullmore Edward, Sporns Olaf. Complex brain networks: Graph theoretical analysis of structural and functional systems *Nature reviews. Neuroscience.* 2009;10:186-98.

[4] Yu Zhe, Prado Raquel, Quinlan Erin Burke, Cramer Steven C., Ombao Hernando. Understanding the Impact of Stroke on Brain Motor Function: A Hierarchical Bayesian Approach *Journal of the American Statistical Association.* 2016;111:549-563.

[5] Horwitz Barry. The elusive concept of brain connectivity *Neuroimage.* 2003;19:466–470.

[6] Friston Karl J. Functional and effective connectivity: a review *Brain connectivity.* 2011;1:13–36.

[7] Chen Jingyuan E, Rubinov Mikail, Chang Catie. Methods and Considerations for Dy-

namic Analysis of Functional MR Imaging Data *Neuroimaging clinics of North America.* 2017;27:547—560.

[8] Warnick Ryan, Guindani M., Erhardt Erik B., Allen Elena A., Calhoun Vince D., Vannucci Marina. A Bayesian Approach for Estimating Dynamic Functional Network Connectivity in fMRI Data *Journal of the American Statistical Association.* 2018;113:134 - 151.

[9] Qiu Huitong, Han Fang, Liu Han, Caffo Brian. Joint estimation of multiple graphical models from high dimensional time series *Journal of the Royal Statistical Society. Series B, Statistical Methodology.* 2016;78:487.

[10] Zhang Aiying, Cai Biao, Hu Wenxing, et al. Joint Bayesian-Incorporating Estimation of Multiple Gaussian Graphical Models to Study Brain Connectivity Development in Adolescence *IEEE Transactions on Medical Imaging.* 2020;39:357 - 365.

[11] Monti Ricardo Pio, Hellyer Peter, Sharp David, Leech Robert, Anagnostopoulos Christoforos, Montana Giovanni. Estimating time-varying brain connectivity networks from functional MRI time series *NeuroImage.* 2014;103:427–443.

[12] Danaher Peter J., Wang Pei, Witten Daniela M.. The joint graphical lasso for inverse covariance estimation across multiple classes *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2011;76.

[13] Peterson Christine B., Stingo Francesco C., Vannucci Marina. Bayesian Inference of Multiple Gaussian Graphical Models *Journal of the American Statistical Association.* 2015;110:159 - 174.

[14] Zhang Linlin, Guindani Michele, Vannucci Marina. Bayesian models for functional magnetic resonance imaging data analysis *Wiley Interdisciplinary Reviews: Computational Statistics.* 2015;7:21-41.

[15] Castruccio Stefano, Ombao Hernando, Genton Marc G.. A scalable multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data *Biometrics.* 2018;74:823-833.

[16] Rajapakse Jagath C, Zhou Juan. Learning effective brain connectivity with dynamic Bayesian networks *NeuroImage.* 2007;37:749—760.

[17] Mumford Jeanette A., Ramsey Joseph. Bayesian networks for fMRI: A primer *NeuroImage.* 2014;86:573-582.

[18] Samdin S., Ting Chee-Ming, Ombao Hernando, Salleh Shussain. A Unified Estimation Framework for State-Related Changes in Effective Brain Connectivity *IEEE Transactions on Biomedical Engineering.* 2016;64.

[19] Ge Bao, Wang Huan, Wang Panpan, Tian Yin, Zhang Xin, Liu Tianming. Discovering and characterizing dynamic functional brain networks in task FMRI *Brain Imaging and Behavior.* 2020;14.

[20] Andersen Michael Riis, Winther Ole, Hansen Lars Kai, Poldrack Russell A., Koyejo Oluwasanmi. Bayesian Structure Learning for Dynamic Brain Connectivity in *AISTATS* 2018.

[21] Liang F., Xue J., Jia B.. Markov Neighborhood Regression for High-Dimensional Inference *Journal of the American Statistical Association.* 2022;117:1200-1214.

[22] Jia Bochao, Liang Faming, TEDDY Study Group . Fast hybrid Bayesian integrative learning of multiple gene regulatory networks for type 1 diabetes *Biostatistics.* 2021;22:233-249.

[23] Liang F., Song Q., Qiu P.. An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models *Journal of the American Statistical Association.* 2015;110:1248-1265.

[24] Hariri Ahmad R, Brown Sarah M, Williamson Douglas E, Flory Janine D, De Wit Harriet, Manuck Stephen B. Preference for immediate over delayed rewards is associated with magnitude of ventral striatal activity *Journal of Neuroscience.* 2006;26:13213–13217.

[25] Geer Sara, Bühlmann Peter, Ritov Y., Dezeure R.. On asymptotically optimal confidence regions and tests for high-dimensional models *Ann. Statist..* 2014;42:1166-1202.

[26] Zhang Cun-Hui, Zhang Stephanie S.. Confidence intervals for low dimensional parameters in high dimensional linear models *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2014;76:217–242.

[27] Javanmard Adel, Montanari Andrea. Confidence intervals and hypothesis testing for high-dimensional regression *Journal of Machine Learning Research.* 2014;15:2869-2909.

[28] Lauritzen S.. *Graphical Models*. Oxford University Press 1996.

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

[30] Zhang Cun-Hui. Nearly unbiased variable selection under minimax concave penalty *Annals of Statistics.* 2010:894–942.

[31] Tibshirani Robert. Regression shrinkage and selection via the lasso *Journal of the Royal Statistical Society. Series B (Methodological).* 1996:267–288.

[32] Bhlmann Peter, Geer Sara A.. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer 2011.

[33] Fan Jianqing, Lv Jinchi. Sure independence screening for ultrahigh dimensional feature space *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2008;70:849–911.

[34] Meinshausen Nicolai, Buhlmann Peter. High-dimensional graphs and variable selection with the Lasso *Ann. Statist..* 2006;34:1436–1462.

[35] Friedman Jerome, Hastie Trevor, Tibshirani Robert. Sparse inverse covariance estimation with the graphical lasso *Biostatistics.* 2008;9:432–441.

[36] Sun Lizhe, Liang Faming. Markov neighborhood regression for statistical inference of high-dimensional generalized linear models *Statistics in Medicine.* 2022;41:4057-4078.

[37] Friston K.J., Harrison L., Penny W.. Dynamic causal modelling *NeuroImage.* 2003;19:1273-1302.

[38] Ryali Srikanth, Supekar Kaustubh, Chen Tianwen, Menon Vinod. Multivariate dynamical systems models for estimating causal interactions in fMRI *NeuroImage.* 2010;54:807-23.

[39] Markett Sebastian A., Jawinski Philippe, Kirsch Peter, Gerchen Martin Fungisai. Specific and segregated changes to the functional connectome evoked by the processing of emotional faces: A task-based connectome study *Scientific Reports.* 2020;10.

[40] Friston Karl J.. Functional and effective connectivity in neuroimaging: A synthesis *Human Brain Mapping.* 1994;2:56-78.

[41] Friedman N., Linial M., Nachman I., Pe'er D.. Using Bayesian networks to analyze expression data *Journal of Computational Biology.* 2000;7:601-620.

[42] Ellis Byron, Wong Wing Hung. Learning Causal Bayesian Network Structures From Experimental Data *Journal of the American Statistical Association.* 2008;103:778-789.

[43] Liang Faming, Zhang Jian. Learning Bayesian networks for discrete data *Computational Statistics & Data Analysis.* 2009;53:865-876.

[44] Meek Christopher. Strong completeness and faithfulness in Bayesian networks in *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995* (Besnard Philippe, Hanks Steve. , eds.):411–418Morgan Kaufmann 1995.

[45] Celeux G., Chauveau D., Diebolt J.. Stochastic versions of the EM algorithm: an experimental study in the mixture case *Journal of Statistical Computation and Simulation.* 1996;55:287-314.

[46] Nielsen Søren. The stochastic EM algorithm: estimation and asymptotic results *Bernoulli.* 2000;6:457-489.

[47] Pellet Jean-Philippe, Elisseeff André. Using Markov Blankets for Causal Structure Learning *J. Mach. Learn. Res..* 2008;9:1295-1342.

[48] Xu Suwa, Jia Bochao, Liang Faming. Learning moral graphs in construction of high-dimensional bayesian networks for mixed data *Neural computation.* 2019;31:1183–1214.

[49] Fan Jianqing, Song Rui. Sure independence screening in generalized linear models with NP-dimensionality *The Annals of Statistics.* 2010;38:3567–3604.

[50] Liang Faming, Jia Bochao. *Sparse Graphical Modeling for High Dimensional Data*. New York: CRC Press 2023.

[51] Saldana Diego Franco, Feng Yang. SIS: An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models *Journal of Statistical Software.* 2018;83:1–25.

[52] Jia Bochao, Liang Faming. Joint estimation of multiple mixed graphical models for pan-cancer network analysis *Stat.* 2020;9:e271.

[53] Liang Faming, Jia Bochao, Xue Jingnan, Li Qizhai, Luo Ye. An imputation–regularized optimization algorithm for high dimensional missing data problems and beyond *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2018;80:899–926.

[54] Zaykin Dmitri V.. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis *Journal of Evolutionary Biology.* 2011;24.

[55] Stouffer S., Suchman E., DeVinney Leland C., Star Shirley A., Jr. Robin M. Williams. *The American Soldier: Adjustment During Army Life. Vol. 1.* Princeton: Princeton University Press 1949.

[56] Owen Art B.. Karl Pearson's meta analysis revisited *Quality Engineering.* 2009;55:493-494.

[57] Liang Faming, Zhang Jian. Estimating the false discovery rate using the stochastic approximation algorithm *Biometrika.* 2008;95:961–977.

[58] Efron Bradley. Large-Scale Simultaneous Hypothesis Testing *Journal of the American Statistical Association.* 2004;99:104 - 96.

[59] Efron Bradley, Tibshirani Robert. Empirical bayes methods and false discovery rates for microarrays *Genetic Epidemiology.* 2002;23.

[60] Efron Bradley, Tibshirani Robert, Storey John D., Tusher Virginia Goss. Empirical Bayes Analysis of a Microarray Experiment *Journal of the American Statistical Association.* 2001;96:1151 - 1160.

[61] Storey J.D.. A direct approach to false discovery rates *Journal of the Royal Statistical Society, Series B.* 2002;64:479-498.

[62] Ha Min, Stingo Francesco, Baladandayuthapani Veerabhadran. Bayesian Structure Learning in Multi-layered Genomic Networks *Journal of the American Statistical Association.* 2020;116:1-33.

[63] Zou Hui, Hastie Trevor. Regularization and variable selection via the elastic net *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2005;67:301–320.

[64] Van Essen David C, Smith Stephen M, Barch Deanna M, et al. The WU-Minn human connectome project: an overview *Neuroimage.* 2013;80:62–79.

[65] Glasser Matthew F, Sotiropoulos Stamatios N, Wilson J Anthony, et al. The minimal preprocessing pipelines for the Human Connectome Project *Neuroimage.* 2013;80:105–124.

[66] Finn Emily S, Shen Xilin, Scheinost Dustin, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity *Nature neuroscience.* 2015;18:1664–1671.

[67] Cai Biao, Zhang Gemeng, Zhang Aiying, et al. Functional connectome fingerprinting: identifying individuals and predicting cognitive functions via autoencoder *Human Brain Mapping.* 2021;42:2691–2705.

[68] Zigmond Michael J, Bloom Floyd E, Landis Story C, Roberts James L, Squire Larry R. *Fundamental neuroscience*;207. Academic press San Diego 1999.

[69] Rajmohan V, Mohandas E. The limbic system *Indian journal of psychiatry.* 2007;49:132.

[70] Ferrucci Roberta, Giannicola Gaia, Rosa Manuela, et al. Cerebellum and processing of negative facial emotions: cerebellar transcranial DC stimulation specifically enhances the emotional recognition of facial anger and sadness *Cognition & emotion.* 2012;26:786–799.

[71] Baumann Oliver, Mattingley Jason B.. Functional topography of primary emotion processing in the human cerebellum *NeuroImage.* 2012;61:805-811.

[72] Adamaszek Michael, D'Agata Federico, Ferrucci Roberta, et al. Consensus Paper: Cerebellum and Emotion *The Cerebellum.* 2017;16:552-576.

[73] Gold Alexandra K., Toomey Rosemary. The role of cerebellar impairment in emotion processing: a case study *Cerebellum & Ataxias.* 2018;5.

[74] Esslen Michaela, Pascual-Marqui Roberto D., Hell Daniel, Kochi Kieko, Lehmann Dietrich. Brain areas and time course of emotional processing *NeuroImage.* 2004;21:1189-1203.

[75] Stuhrmann Anja, Suslow Thomas, Dannlowski Udo. Facial emotion processing in major depression: a systematic review of neuroimaging findings *Biology of Mood & Anxiety Disorders.* 2011;1:10 - 10.

Table 1: Comparison of JGGM+MNR and BANS for the simulated examples with the sample size $n = 400$, varying numbers of nodes ($p = 20$, 30, 50), and the number of time points $T = 60$. † Average AUC values produced with hyper-parameters $(a_1, b_1) = (3, 1)$, $(5, 1)$, and $(10, 1)$; ‡ average computational time measured in minutes on a computer with i9-10900K CPU@3.6GHz and 128GB of memory, running in parallel with 20 threads.

		JGGM+MNR			BANS
	Parameter	$a_1 = 3, b_1 = 1$	$a_1 = 5, b_1 = 1$	$a_1 = 10, b_1 = 1$	-
$p = 20$	AUC†	0.959(0.007)	0.959(0.007)	0.959(0.007)	0.839(0.008)
	Time‡	5.18	5.064	5.268	47.93
$p = 30$	AUC†	0.952(0.005)	0.952(0.004)	0.952(0.005)	0.766(0.006)
	Time‡	5.891	5.786	5.792	96.22
$p = 50$	AUC†	0.944(0.004)	0.945(0.003)	0.945(0.004)	-
	Time‡	7.478	7.723	7.676	$>$ 12 hours

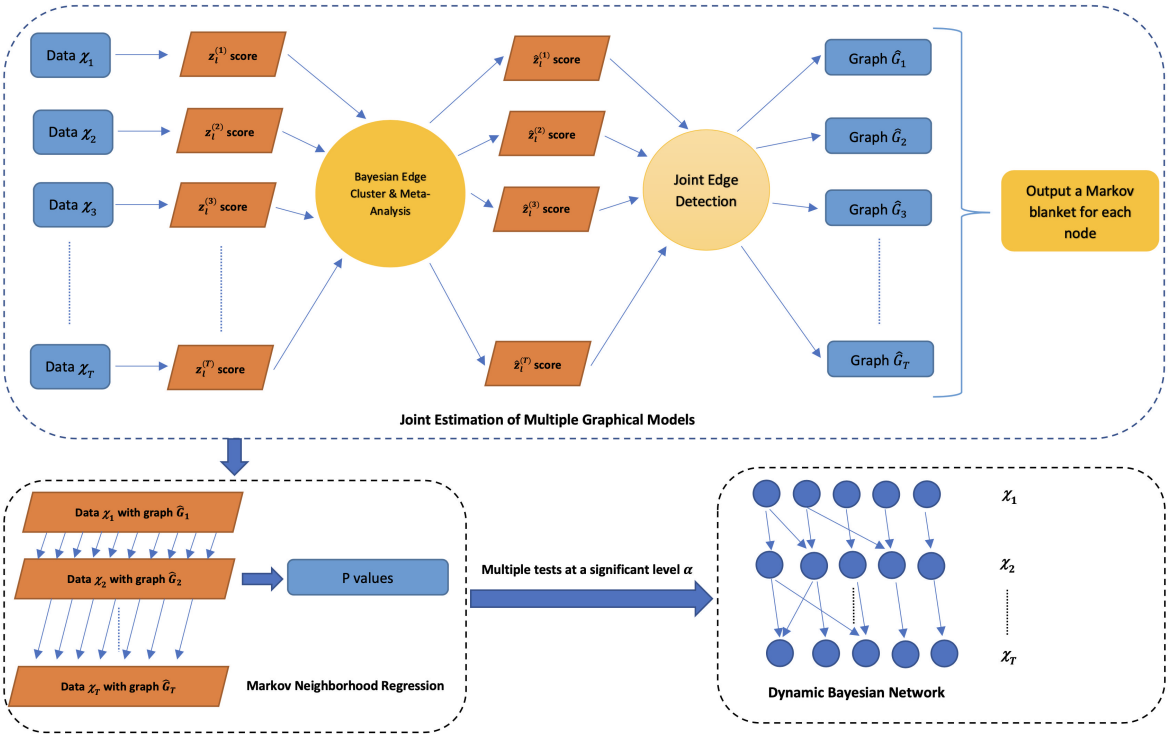

Figure 1: Workflow of the proposed method for construction of a dynamic Bayesian network: The upper block represents the first step, which outputs a Markov blanket for each node at each time point; the lower two blocks represent the second step, where a p-value is calculated for each possible edge of the dynamic Bayesian network using Markov neighborhood regression and, subsequently, a multiple hypothesis test is performed on the p-values to identify the significant edges of the dynamic Bayesian network.

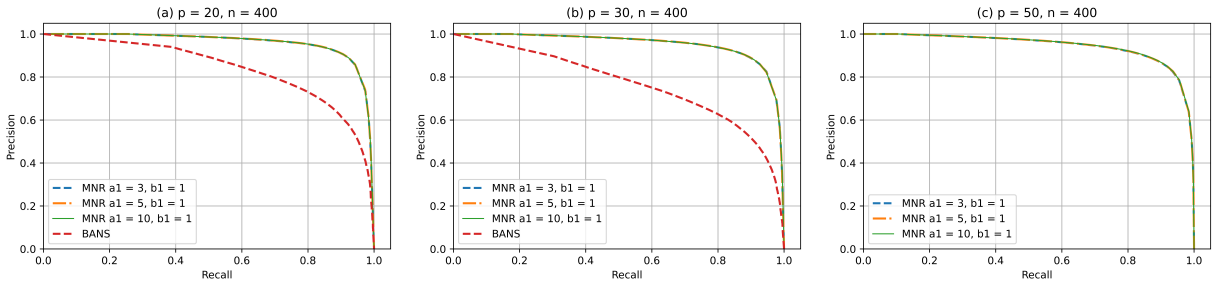

Figure 2: Precision-recall curves produced by JGGM+MNR and BANS for the simulated examples under the settings: (a) $p = 20$, $n = 400$, $T = 60$; (b) $p = 30$, $n = 400$, $T = 60$; (c) $p = 50$, $n = 400$, $T = 60$, for which BANS is time-consuming (with computational time > 12 hours) and thus omitted for comparison.

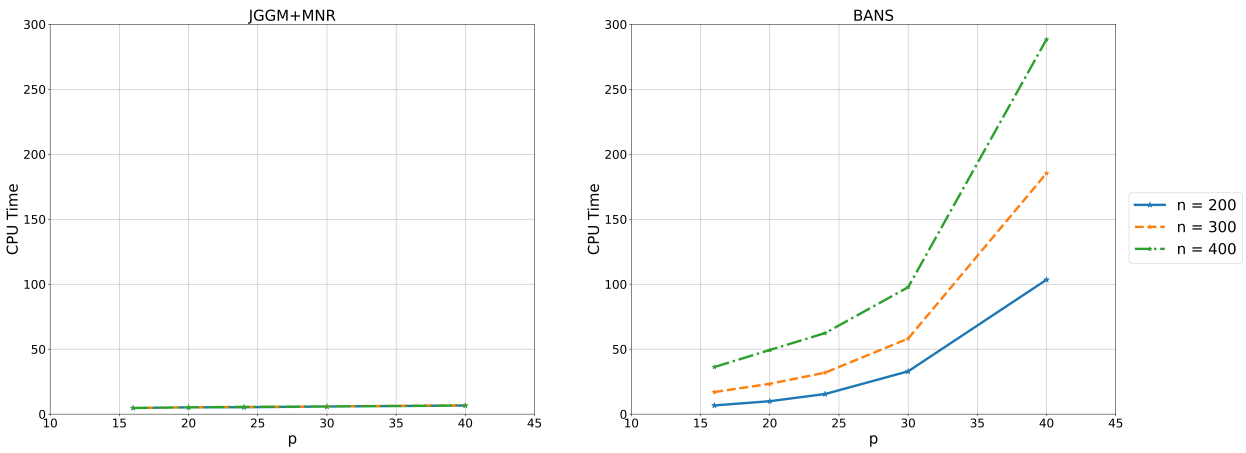

Figure 3: Time complexity curves of JGGM+MNR and BANS for the simulated examples with varying sample sizes $n = 200, 300, 400$, varying numbers of nodes $p = 16, 20, 24, 30, 40$, and the number of time points $T = 60$.

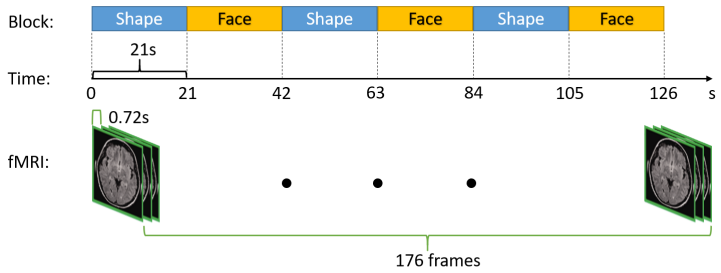

Figure 4: Illustration of the fMRI experimental design under the emotion processing task.

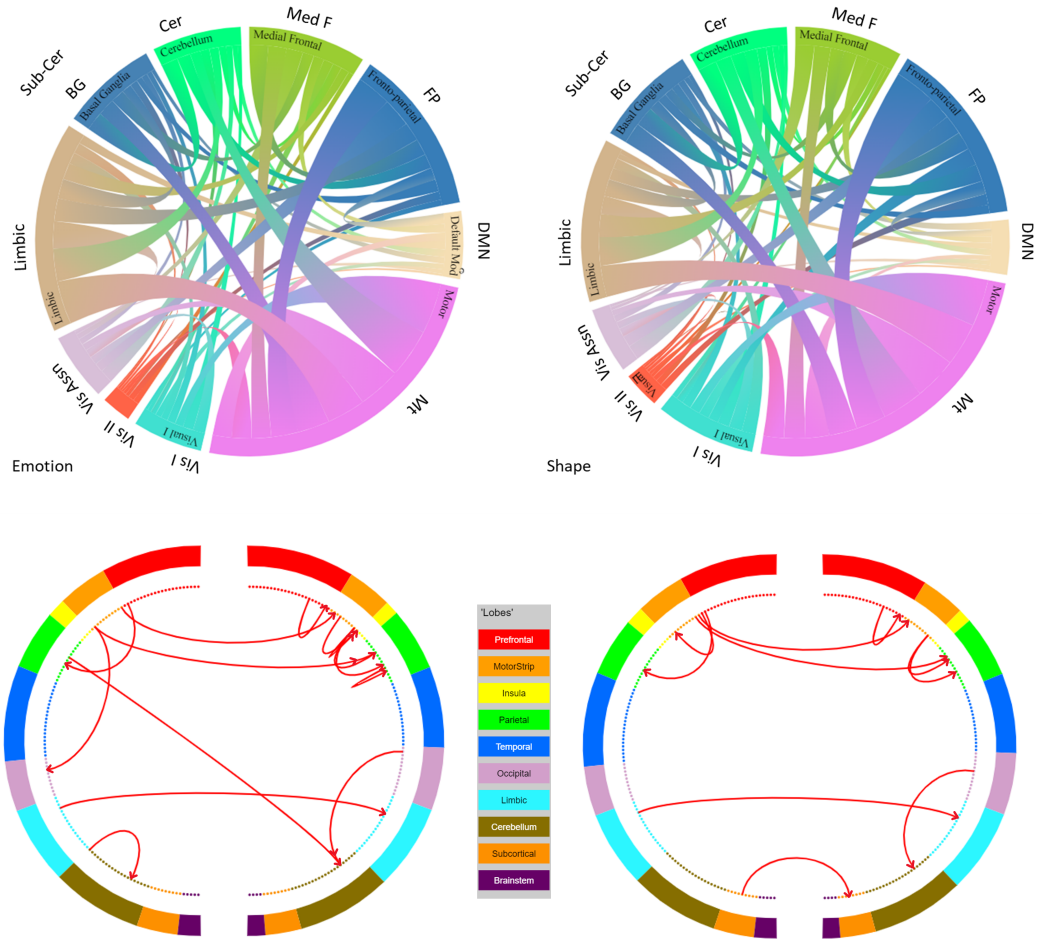

Figure 5: Chord plots arranged by eight functional network modules for emotion processing (upper left) and shape processing (upper right), and characteristic edges for emotion processing (down left) and shape processing (down right).

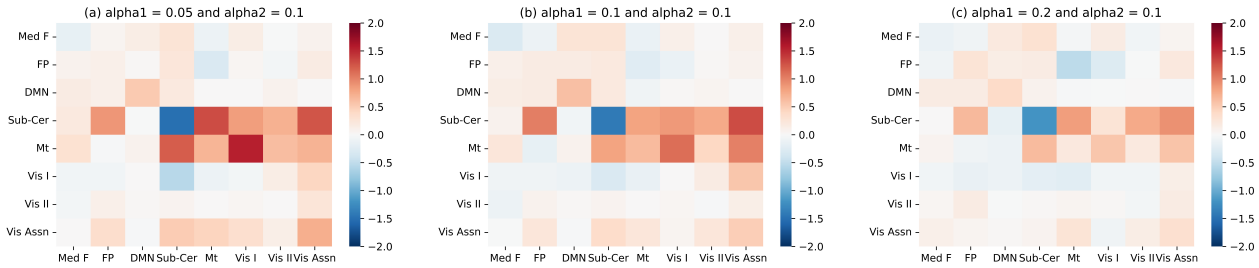

Figure 6: Heat maps for the mean edge (functional module-wise) difference between the emotion and shape tasks, identified under the settings (a) $(\alpha_1, \alpha_2) = (0.05, 0.1)$, (b) $(\alpha_1, \alpha_2) = (0.1, 0.1)$, and (c) $(\alpha_1, \alpha_2) = (0.2, 0.1)$.

Algorithm 2 Edgewise Score Evaluation

Input: The observations in a three-dimensional set $\{\mathcal{X}_1, \mathcal{X}_2, \cdots \mathcal{X}_T\}$, where T denotes the total number of different time points and each $\mathcal{X}_t$ is $n_t \times p$ data matrix.

Step 1. (**Super-Markov Blankets Construction**) For each variable, treat it as the response variable to conduct sure independence screening[33] with respect to the remaining variables, and limit its super-Markov blanket size to the order $O(n/\log(n))$. In practice, the super-Markov blanket size is $n/c_n \log(n)$, where c_n is a tunable parameter with a default value of 1.

Step 2. (**Conditional Independence Tests**) For each ordered pair of variables $(X_{t,i}, X_{t,j})$, $i = 1, 2, \cdots p$ and $j = i+1, i+2, \ldots, p$, $t = 1, 2, \cdots, T$, conduct the conditional independence test:

$$X_{t,i} \perp X_{t,j} \mid \tilde{\boldsymbol{S}}_{t,ij} \backslash \{X_{t,i}, X_{t,j}\},\ i, j = 1, 2, \cdots, p,\ t = 1, 2, \cdots, T,$$

where $\tilde{\boldsymbol{S}}_{t,ij} = \tilde{\boldsymbol{S}}_{t,i}$ if $|\tilde{\boldsymbol{S}}_{t,i} \backslash \{X_{t,i}, X_{t,j}\}| \leq |\tilde{\boldsymbol{S}}_{t,j} \backslash \{X_{t,i}, X_{t,j}\}|$ and $\tilde{\boldsymbol{S}}_{t,ij} = \tilde{\boldsymbol{S}}_{t,j}$ otherwise, and denote the p-values of the test by $p_{ij}^{(t)}$. For multivariate Gaussian data, the relationships among conditional distributions, partial correlation coefficients, and precision matrix, see Liang and Jia[50] (p.2) or Bühlmann and van de Geer[32] (p.436), this test can be performed by regressing $X_{t,i} = \beta_{t,j} X_{t,j} + \boldsymbol{\beta}_{t,\tilde{\boldsymbol{S}}} \tilde{\boldsymbol{S}}_{t,ij} + \epsilon_{t,i}$ and then testing the hypothesis $H_0 : \beta_{t,j} = 0$ versus $H_1 : \beta_{t,j} \neq 0$.

Step 3. (z**-scores Calculation**) Convert the p-values of the conditional independence tests to z-scores by the probit transformation

$$z_{ij}^{(t)} = \Phi^{-1}(1 - p_{ij}^{(t)}),\ t = 1, 2, \cdots, T,$$

where Φ^{-1} is the Gaussian inverse transformation and $p_{ij}^{(t)}$ is the p-value of the conditional independence test for the pair $(X_{t,i}, X_{t,j})$.

Output: The z-scores for each ordered pair of variables $(X_{t,i}, X_{t,j})$ of each data set $\mathcal{X}_t$, where $t = 1, 2, \cdots, T$.

Algorithm 3 Bayesian Edge Clustering

Input: Randomly initialize $e_{l,0}^{(1)}$ in $\{0,1\}$, and then draw $e_{l,0}^{(t)}$ according to the Bernoulli distribution $P(e_{l,0}^{(t)} = e_{l,0}^{(t-1)}|e_{l,0}^{(t-1)}) = q$ and $P(e_{l,0}^{(t)} = 1-e_{l,0}^{(t-1)}|e_{l,0}^{(t-1)}) = 1-q$, $t = 1, 2, \cdots, T$.

for $l = 1, 2, \ldots, N$ **do**

for $k = 0, 1, \ldots, \mathcal{K}$ **do**

Optimization: Cluster $z_l^{(1)}, z_l^{(2)}, \ldots, z_l^{(T)}$ according to $\boldsymbol{e}_{l,k}$; estimate the mean and variance of each cluster, denoting the respective estimates by $(\hat{\mu}_{l0}^{(k)}, \hat{\sigma}_{l0}^{(k)})$ and $(\hat{\mu}_{l1}^{(k)}, \hat{\sigma}_{l1}^{(k)})$; find the values of $n_1 = |\{t : e_{l,k}^{(t)} = e_{l,k}^{(t-1)}\}|$ and $T - n_1 - 1 = |\{t : e_{l,k}^{(t)} = 1 - e_{l,k}^{(t-1)}\}|$, and estimate q by the mean of Beta$(a_1 + n_1, b_1 + T - n_1 - 1)$.

Imputation: Update $\boldsymbol{e}_{l,k}$ by using the Gibbs sampler in the following steps.

Step 1. For $t = 1$, calculate

$$S_{1,k}^{(0)} = \mathcal{N}(z_l^{(1)}; \hat{\mu}_{l0}^{(k)}, \hat{\sigma}_{l0}^{(k)}) \times q^{1-c_{1,k}^{(0)}}(1-q)^{c_{1,k}^{(0)}},$$
$$S_{1,k}^{(1)} = \mathcal{N}(z_l^{(1)}; \hat{\mu}_{l1}^{(k)}, \hat{\sigma}_{l1}^{(k)}) \times q^{1-c_{1,k}^{(1)}}(1-q)^{c_{1,k}^{(1)}},$$

where $c_{1,k}^{(0)} = |e_{l,k}^{(2)} - 0|$ and $c_{1,k}^{(1)} = |e_{l,k}^{(2)} - 1|$.

Step 2. For $t = 2, 3, \cdots, T-1$, calculate

$$S_{t,k}^{(0)} = \mathcal{N}(z_l^{(t)}; \hat{\mu}_{l0}^{(k)}, \hat{\sigma}_{l0}^{(k)}) \times q^{2-(c_{t-1,k}^{(0)} + c_{t,k}^{(0)})}(1-q)^{c_{t-1,k}^{(0)} + c_{t,k}^{(0)}},$$
$$S_{t,k}^{(1)} = \mathcal{N}(z_l^{(t)}; \hat{\mu}_{l1}^{(k)}, \hat{\sigma}_{l1}^{(k)}) \times q^{2-(c_{t-1,k}^{(1)} + c_{t,k}^{(1)})}(1-q)^{c_{t-1,k}^{(1)} + c_{t,k}^{(1)}},$$

where $c_{t-1,k}^{(0)} = |0 - e_{l,k}^{(t-1)}|$, $c_{t-1,k}^{(1)} = |1 - e_{l,k}^{(t-1)}|$, $c_{t,k}^{(0)} = |e_{l,k}^{(t+1)} - 0|$, and $c_{t,k}^{(1)} = |e_{l,k}^{(t+1)} - 1|$.

Step 3. For t = T, calculate

$$S_{T,k}^{(0)} = \mathcal{N}(z_l^{(T)}; \hat{\mu}_{l0}^{(k)}, \hat{\sigma}_{l0}^{(k)}) \times q^{1-c_{T,k}^{(0)}}(1-q)^{c_{T,k}^{(0)}},$$
$$S_{T,k}^{(1)} = \mathcal{N}(z_l^{(T)}; \hat{\mu}_{l1}^{(k)}, \hat{\sigma}_{l1}^{(k)}) \times q^{1-c_{T,k}^{(1)}}(1-q)^{c_{T,k}^{(1)}},$$

where $c_{T,k}^{(0)} = |0 - e_{l,k}^{(t-1)}|$ and $c_{T,k}^{(1)} = |1 - e_{l,k}^{(t-1)}|$.

Step 4. Draw $e_{l,k+1}^{(t)}$ for $t = 1, 2, \ldots, T$ according to the distribution:

$$\mathbb{P}(e_{l,k+1}^{(t)} = 0|e_{l,k}^{(t-1)}, e_{l,k}^{(t+1)}) = \frac{S_{t,k}^{(0)}}{S_{t,k}^{(0)} + S_{t,k}^{(1)}}, \quad \mathbb{P}(e_{l,k+1}^{(t)} = 1|e_{l,k}^{(t-1)}, e_{l,k}^{(t+1)}) = \frac{S_{t,k}^{(1)}}{S_{t,k}^{(0)} + S_{t,k}^{(1)}}.$$

end for

end for

Output: Estimates for the status of the edges $\{\mathbf{e}_l : l = 1, 2, \cdots, N\}$.

Algorithm 4 Time-Varying Dynamic Bayesian Network Learning

Input: The dataset $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T\}$, where $\mathcal{X}_t$ is an $n_t \times p$-matrix and each row of $\mathcal{X}_t$ follows a multivariate Gaussian distribution $N_p(0, \Sigma_t)$ for $t = 1, 2, \ldots, T$.

1. **Joint estimation of multiple graphical models.** Apply the accelerated hybrid Bayesian integrative analysis method to the data set to jointly estimate T Gaussian graphical models. Denote the graphical models by $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T$.

2. **Markov neighborhood regression.** For $t = 2, 3, \ldots, T$ and $j = 1, 2, \ldots, p$, conduct statistical inference for the high-dimensional regression

$$X_{t,j} = \beta_{t,j}^{(0)} + \beta_{t,j}^{(1)} X_{t-1,1} + \beta_{t,j}^{(2)} X_{t-1,2} + \cdots + \beta_{t,j}^{(p)} X_{t-1,p} + \epsilon_{t,j}, \quad \epsilon_{t,j} \sim N(0, \sigma_{t,j}^2), \tag{14}$$

   using the MNR method for which the Markov blanket estimate is obtained from $\mathcal{G}_{t-1}$; and denote the p-values corresponding to the variables $X_{t-1,1}, X_{t-1,2}, \ldots, X_{t-1,p}$ by $q_{1,j}^{(t)}, q_{2,j}^{(t)}, \ldots, q_{p,j}^{(t)}$, respectively.

3. **Joint Link Detection.** Transform the p-values obtained in the MNR step to z-scores by setting

$$z_{i,j}^{(t)} = \Phi^{-1}(1 - q_{i,j}^{(t)}), \quad i, j = 1, 2, \ldots, p, \text{ and } t = 2, 3, \ldots, T,$$

   and conduct multiple hypothesis tests for all $z_{i,j}^{(t)}$'s to detect the values that are significantly larger than others. The p-values will be adjusted by using the empirical Bayesian method.[57] Set the elements of the adjacency matrix of the dynamic Bayesian network to 1 if the corresponding z-scores are significantly larger than others and 0 otherwise.

Output: The adjacency matrix of the dynamic Bayesian network from $t-1$ to t, where $t = 2, 3, \cdots, T$.

The Supplement for "Time-Varying Dynamic Bayesian Network Learning for an fMRI Study of Emotion Processing"

Lizhe Sun, Aiying Zhang and Faming Liang *

§1 Theoretical Justification

The consistency of Algorithm 4 for dynamic Bayesian network learning can be justified based on the consistency of the MNR method for the identification of the directed structure around the response variable, as established in Liang et al.[1] As implied by equations (6)-(8) in the main text, we require the conditions that ensure the consistency of high-dimensional variable selection for each node at each time point, the conditions that ensure the consistency of the joint estimates of the multiple Gaussian graphical models, as well as the conditions that guarantee the sparsity of the dynamic Bayesian network and the multiple Gaussian graphical models. The conditions for the consistency of the joint estimation of multiple Gaussian graphical models and their sparsity have been specified in Jia et al.[2] The conditions for the consistency of variable selection in each MNR and its sparsity have been provided in Liang et al.[1] To ensure the paper self-contained, we summarize these conditions and provide a brief justification for the consistency of Algorithm 4 in the following. To indicate the dependency of the dimension p on the sample size n, we denote it by p_n.

*To whom correspondence should be addressed: F. Liang. Liang is Distinguished Professor (email: fmliang@purdue.edu), Department of Statistics, Purdue University, West Lafayette, IN 47907. Sun is Postdoc, Beijing International Center for Mathematical Research, Peking University and Department of Statistics, Purdue University. Zhang is Assistant Professor, School of Data Science, University of Virginia.

(A0) Let $n = \min\{n_1, n_2, \cdots, n_T\}$; and let the dimension $p_n = O(\exp(n^\delta))$ for some constant $0 \leq \delta < 1/2$.

(A1) For any $t \in \{1, 2, \ldots, T\}$, $\boldsymbol{X}_t$ satisfies the following conditions:

(i) The joint distribution $P(\boldsymbol{X}_t)$ is multivariate Gaussian, and it satisfies the Markov property and adjacency faithfulness condition with respect to the undirected underlying graph $\mathcal{G}(\boldsymbol{X}_t)$.

(ii) The correlation coefficients satisfy $\min\{|r_{ij}|; e_{ij} = 1, i, j = 1, \ldots, p_n, i \neq j\} \geq c_0 n^{-\kappa}$ for some constants $c_0 > 0$ and $0 < \kappa < (1-\delta)/2$, and $\max\{|r_{ij}|; i, j = 1, \ldots, p_n, i \neq j\} \leq M_r < 1$ for some constants $0 < M_r < 1$.

(iii) There exists constants $c_1 > 0$, $0 < \kappa' \leq \kappa$, and $0 \leq \tau < 1 - 2\kappa'$ such that $\lambda_{\max}(\Sigma_{x,t}) \leq c_1 n^\tau$, where $\Sigma_{x,t}$ denotes the covariance matrix of $\boldsymbol{X}_t$.

(iv) The ψ-partial correlation coefficients, as defined in Liang et al.[3], satisfy $\inf\{\psi_{ij} : \psi_{ij} \neq 0, i, j = 1, \ldots, p_n, i \neq j, |S_{ij}| \leq q_n\} \geq c_2 n^{-d}$, where $q_n = O(n^{2\kappa'+\tau}), 0 < c_2 < \infty$, $0 < d < (1-\delta)/2$ are some constants, and S_{ij} denotes the conditioning set used in calculating ψ_{ij}. In addition, $\sup\{\psi_{ij} : i, j = 1, \ldots, p_n, i \neq j, |S_{ij}| \leq q_n\} \geq c_6 n^{-d} \leq M_\psi < 1$ for some constants $0 < M_\psi < 1$.

(A2) $\max_{t\in\{1,2,\ldots,T\}, j\in\{1,\ldots,p_n\}} |\xi_{t,j}| = o(n^{1/2})$, where $\xi_{t,j}$ denotes the Markov blanket of node j at time t with respect to the Gaussian graphical model formed by the data $\mathcal{X}_t$.

(A3) There exist constants $c_3 > 0$ and $c_4 > 0$ such that $\min_{t\in\{2,\ldots,T\}, j\in\{1,2,\ldots,p_n\}, i\in \boldsymbol{S}^*_{t,j}} |\beta^{(i)}_{t,j}| \geq c_3 n^{-\kappa}$ and $\min_{t\in\{2,\ldots,T\}, j\in\{1,2,\ldots,p_n\}, i\in \boldsymbol{S}^*_{t,j}} |cov((\beta^{(i)}_{t,j})^{-1} X_{t,j}, X_{t-1,i})| \geq c_4$, where $\boldsymbol{S}^*_{t,j} = \{i : \beta^{(i)}_{t,j} \neq 0, i \in \{1, 2, \ldots, p_n\}\}$ denotes the set of true variables of the linear regression model

$$X_{t,j} = \beta^{(0)}_{t,j} + \beta^{(1)}_{t,j} X_{t-1,1} + \beta^{(2)}_{t,j} X_{t-1,2} + \cdots + \beta^{(p_n)}_{t,j} X_{t-1,p_n} + \epsilon_{t,j}, \tag{S1}$$

and $\epsilon_{t,j} \sim N(0, \sigma^2_{t,j})$ denotes a zero-mean Gaussian random error.

(A4) $\max_{t\in\{2,\ldots,T\}, j\in\{1,2,\ldots,p_n\}} |\boldsymbol{S}^*_{t,j}| = o(n^{1/3})$.

(A5) Other assumptions in Theorem 2 of Fan and Peng[4] for each regression (S1) with $t \in \{2, \ldots, T\}$ and $j \in \{1, 2, \ldots, p_n\}$.

(A6) Other assumptions in Theorem 1 of Fan et al[5] (the case of random design) for each regression (S1) with $t \in \{2, \ldots, T\}$ and $j \in \{1, 2, \ldots, p_n\}$.

(A7) The number of distinct conditions $T_n = O(n^{\delta+2d+\varepsilon-1})$ for some constant $\varepsilon > 0$ such that $\delta + 2d + \varepsilon - 1 \geq 0$, where δ is as defined in (A0) and d is as defined in (A1)-(iv).

Theorem S1. *Let $n = \min\{n_1, n_2, \cdots, n_T\}$. Suppose that conditions (A0)-(A7) (given in the supplementary material) hold. Then the dynamic Bayesian network estimator resulting from Algorithm 4 is consistent.*

Proof. With conditions (A0), (A1), and (A7), we can apply Theorem 1 of Jia and Liang[2] to prove that $P(\hat{\xi}_{t,j} = \xi_{t,j}) = 1 - o(1)$ for all $j = 1, \ldots, p_n$ and $t = 1, 2, \ldots, T$. With conditions (A0), (A1)-(iii) and (A3)-(A5), we can apply Theorem 5 of Fan and Lv[6] to prove that

$$P(\hat{\boldsymbol{S}}^*_{t,j} = \boldsymbol{S}^*_{t,j}) = 1 - o(1),\ j = 1, 2, \cdots, p_n,\ t = 2, \cdots T,\ \text{as } n \to \infty.$$

where $\hat{\boldsymbol{S}}^*_{t,j}$ denotes the estimator of $\boldsymbol{S}^*_{t,j}$ obtained by using the SIS-SCAD method.[6]

With conditions (A2) and (A4), for any $t \in \{2, 3, \ldots, T\}$ and any predictors $i, j \in \{1, 2, \ldots, p_n\}$, we have $D^{(i)}_{t,j} = \{i\} \cup \hat{\xi}_{t-1,i} \cup \hat{\boldsymbol{S}}^*_{t,j}$ and $P(|D^{(i)}_{t,j}| = |\{i\} \cup \hat{\xi}_{t-1,i} \cup \hat{\boldsymbol{S}}^*_{t,j}| = o(n^{1/2})) = 1 - o(1)$, where $D^{(i)}_{t,j}$ denotes the Markov neighborhood of predictor i in the regression (S1). Following from Lemma 1 of Liang et al,[1] we can derive the asymptotic distribution of $\hat{\beta}^{(i)}_{t,j}$. By condition (A6), we can get the consistency of $\hat{\sigma}^2_{t,j}$ based on Theorem 1 of Fan et al[5] and the asymptotic $P(|D^{(i)}_{t,j}| \log(p_n) = o(n)) = 1 - o(1)$ for any $t \in \{2, \ldots, T\}$ and $j \in \{1, 2, \ldots, p_n\}$. Finally, based on Slutsky's theorem, we can conclude that $\sqrt{n}\frac{\hat{\beta}^{(i)}_{t,j} - \beta^{(i)}_{t,j}}{\sqrt{\hat{\sigma}^2_{t,j}\hat{\theta}_{t,jj}}} \sim N(0, 1)$ for any $t \in \{2, \ldots, T\}$, $j \in \{1, 2, \ldots, p_n\}$ and $i \in \{1, 2, \ldots, p_n\}$, where $\hat{\sigma}^2_{t,j}$ denotes an OLS estimate of $\sigma_{t,j}$ from the regression (S1), $\hat{\theta}_{t,jj}$ is the (j, j)-th entry of the matrix $\left[\frac{1}{n}\sum_{k=1}^{n} \tilde{X}^{(k)}_{D^{(i)}_{t,j}} (\tilde{X}^{(k)}_{D^{(i)}_{t,j}})^T\right]^{-1}$, and $\tilde{X}^{(k)}_{D^{(i)}_{t,j}}$ denotes the kth-row of $\boldsymbol{X}_{D^{(i)}_{t,j}}$ and $\boldsymbol{X}_{D^{(i)}_{t,j}}$ denotes a submatrix of $\boldsymbol{X}_{t-1}$ formed with the columns belonging to $D^{(i)}_{t,j}$. Further, following the arguments given in Section 4.1 of Liang et al,[1] we can conclude that the causal structure of the dynamic Bayesian network can be identified by the proposed method. □

§2 More Results on Simulation Studies

§2.1 Data Generation

To mimic the real task-based fMRI data, we generated data from equations (9) to (11) of the main text with a lag order of $L = 1$, where $\gamma_{rk}^{(i)}$ for $r = 1, 2, \ldots, p$ is drawn from the uniform distribution $\mathcal{U}(0, 5)$, $\boldsymbol{\mu}^{(i)} = 1$ for each subject, $\boldsymbol{\epsilon}_t^{(i)} \sim N_p(0, \Sigma)$ with $\Sigma = \text{diag}\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\}$ and each σ_j^2 is drawn from the uniform distribution $\mathcal{U}(0.9, 1.2)$. In this simulation, the canonical HRF $h_r(t)$ is used to model the stimulus, as defined in equation (10) of the main text.

We designed a session comprising two tasks separated by a break. Each task lasted for 25 seconds, while the break lasted for 10 seconds. Consequently, we assigned time points as follows: $t = 1, 2, \cdots, 25$ for task one, $t = 26, 27, \cdots, 35$ for the break, and $t = 36, 37, \cdots, 60$ for task two. In addition, we defined the ROIs $1 : (p/2)$ as blocked module one and ROIs $(p/2+1) : p$ as blocked module two, where p is the number of total ROIs. Each module represents a functional network composed of selected ROIs, which may be specifically activated by certain tasks. In this simulation, we assumed that only module one was activated during the first task, and only module two was activated during the second task. During the break time, neither of the modules was active.

We generated the transition matrix $\boldsymbol{A}_{t,l}$ in the following procedure. First, we generated an initialized $\boldsymbol{A}_{0,l}$ as follows:

$$(\boldsymbol{A}_{0,l})_{i,j} = \begin{cases} 1, & \text{if } |i-j| = 0,\ i = 1, 2, \cdots p, \\ \rho, & \text{if } |i-j| = 1,\ i = 1, 2, \cdots, p-1, \\ \rho^2, & \text{if } |i-j| = 2,\ i = 1, 2, \cdots, p-2, \\ 0, & \text{others,} \end{cases}$$

where $\rho = 0.9$ for the case of strong signal and $\rho = 0.5$ for the case of weak signal. Second, we employed the following procedure to generate the time-varying transition matrix $\boldsymbol{A}_{t,l}$, $t = 1, 2, \cdots, 60$. For task one, we added additional 10% non-zero parameters in the blocked submatrix of initialized transition matrix $\boldsymbol{A}_{0,l}[1 : p/2, 1 : p/2]$. These parameters are drawn uniformly from the interval $[-0.9, -0.6] \bigcup [0.6, 0.9]$ for the case of strong signal

and $[-0.5,-0.3]\bigcup[0.3,0.5]$ for the case of weak signal. During the period of task one, $t=1,2,\cdots,25$, we changed 5% entries to zero and another 5% entries to non-zero in this blocked submatrix of transition matrix $\boldsymbol{A}_{t,l}[1:p/2,1:p/2]$ from t to $t+1$, except diagonal elements. The non-zero elements are drawn uniformly from the interval $[-0.9,-0.6]\bigcup[0.6,0.9]$ for the case of strong signal and $[-0.5,-0.3]\bigcup[0.3,0.5]$ for the case of weak signal. Then, during the resting stage $t=26,27,\cdots 35$, let the transition matrix $\boldsymbol{A}_{t,l}=\boldsymbol{A}_{0,l}$. For task two, we added 10% extra connections in the blocked submatrix of initialized transition matrix $\boldsymbol{A}_{0,l}[(p/2+1):p,(p/2+1):p]$, in which the parameters are drawn uniformly from $[-0.9,-0.6]\bigcup[0.6,0.9]$ for the case of strong signal and $[-0.5,-0.3]\bigcup[0.3,0.5]$ for the case of weak signal. During the period of task two, $t=36,37,\cdots,60$, we changed 5% entries to zero and another 5% entries to non-zero in this blocked submatrix of transition matrix $\boldsymbol{A}_{t,l}[(p/2+1):p,(p/2+1):p]$ from t to $t+1$, except for the diagonal elements. The non-zero elements are drawn uniformly from the interval $[-0.9,-0.6]\bigcup[0.6,0.9]$ for strong signal case and $[-0.5,-0.3]\bigcup[0.3,0.5]$ for weak signal case. All off-diagonal elements are changeable during tasks.

In analyzing the data, we regressed the strength of activation for all nodes and then learned dynamic Bayesian networks from the residual term by using our proposed method, as presented in equation (11) of the main text. The performance of the proposed method is evaluated by using precision-recall curves. We presented AUC values and running time for a sample size of $n=400$ and different numbers of nodes $p=20,30,50$. We fixed the parameter $\alpha_1=0.2$ in all simulations and varied α_2 to construct the precision-recall curve. For the Beta prior of q, we conducted a sensitivity analysis by setting $a_1=3,5,10$ and $b_1=1$. The numerical results of AUC values for strong signals were presented in Table 1 of the main text, and the numerical results of sensitivity analysis were presented in this supplementary material. Each simulation experiment was independently run 10 times (a different dataset was generated each time) on the same personal computer with i9-10900k CPU@3.6GHz and 128G memory, with 20 threads running in parallel.

The precision and recall are defined as follows:

$$\text{precision}=\frac{TP}{TP+FP},\quad \text{recall}=\frac{TP}{TP+FN},$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively, as defined in Table S1.

Table S1: Outcomes of Binary Decision

	True	False
Predicted Positive	True Positive(TP)	False Positive(FP)
Predicted Negative	False Negative(FN)	True Negative(TN)

In the following subsections, we assessed the performance of the proposed method under different scenarios, including large-scale and high-dimensional cases, weak signals, non-stationary connection strengths for each subject, and spatial correlations. In all of these cases, the proposed method consistently outperforms the BANS method.

§2.2 Weak Singals

In this subsection, we assessed the performance of the proposed method in the case of weak signals. In this particular case, we set $\rho = 0.5$, and drew the signal strength parameters uniformly from the interval $[-0.5, -0.3] \cup [0.3, 0.5]$. The results were presented in Table S2.

§2.3 Spatial Correlations

In this subsection, we assessed the performance of the proposed method under the scenario of spatially correlated random errors, i.e., Σ in equation (11) of the main text is non-diagonal.

We generated data under a strong signal setting with $\rho = 0.9$ and the signal strength parameters being drawn uniformly from the interval $[-0.9, -0.6] \bigcup [0.6, 0.9]$. The random noise $\boldsymbol{\epsilon}_t^{(i)}$ in equation (11) of the main text follows a multivariate normal distribution $N_p(0, \Sigma)$,

Table S2: Comparison of JGGM+MNR and BANS for the simulated example under the scenario of weak signals, where the data were simulated with the sample size $n = 400$, varying numbers of nodes ($p = 20, 30, 50$), and the number of time points $T = 60$; the average AUC was produced with the hyper-parameters $(a_1, b_1) = (3, 1)$, $(5, 1)$, and $(10, 1)$; and the average computational time was measured in minutes on a computer with i9-10900K CPU@3.6GHz and 128GB of memory, running in parallel with 20 threads.

		JGGM+MNR			BANS
	Parameter	$a_1 = 3, b_1 = 1$	$a_1 = 5, b_1 = 1$	$a_1 = 10, b_1 = 1$	-
$p = 20$	AUC	0.924(0.005)	0.923(0.005)	0.924(0.005)	0.774(0.005)
	Time	5.286	5.311	5.298	45.19
$p = 30$	AUC	0.898(0.003)	0.898(0.003)	0.898(0.004)	0.691(0.002)
	Time	5.972	5.896	5.905	81.48
$p = 50$	AUC	0.871(0.003)	0.872(0.003)	0.872(0.003)	-
	Time	7.618	7.648	7.281	$>$ 12 hours

and Σ is the covariance matrix with the block structure:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \gamma_{13} & \gamma_{14} & ... & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & ... & 0 & 0 \\ \gamma_{31} & 0 & \sigma_3^2 & \gamma_{34} & ... & 0 & 0 \\ \gamma_{41} & 0 & \gamma_{43} & \sigma_4^2 & ... & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & ... & \sigma_{p-1}^2 & \gamma_{(p-1)p} \\ 0 & 0 & 0 & 0 & ... & \gamma_{p(p-1)} & \sigma_p^2 \end{bmatrix},$$

where each diagonal element σ_j^2 follows the uniform distribution $\mathcal{U}(0.9, 1.2)$, and γ_{ij} appears in the blocked modules $\Sigma[1 : p/2, 1 : p/2]$ and $\Sigma[(p/2 + 1) : p, (p/2 + 1) : p]$ by a completely random pattern. In this experiment, we set $p = 50$. Each γ_{ij} is drawn from a uniform distribution $\mathcal{U}(0, 0.25)$ and $\gamma_{ij} = \gamma_{ji}$ is enforced to ensure that Σ of the main text is symmetric and positive definite. The parameter $\gamma_{ij} = \gamma_{ji} \neq 0$ represents the ROI i and ROI j are spatially correlated. We assume that the ROIs in module one are not spatially correlated

with those in module two. In our real data experiment, the ROIs were generated using the method of Shen et al.[7] Therefore, the spatial correlations among the ROIs are very weak. Additionally, some ROIs might not be spatially correlated, for instance, an ROI in the left hemisphere is not correlated with an ROI in the right hemisphere.

In this simulation study, the average AUC values we presented in Table S3 were calculated based on the simulated dynamic Bayesian networks. Since the undirected edges at time point t were not inferred based on Σ, we only considered the estimation accuracy of the directed edges from nodes $X_{t-1,k}$'s to the nodes $X_{t,j}$'s, where $k, j \in \{1, 2, \ldots, p\}$. See Table S3 for the numerical results.

Table S3: Comparison of JGGM+MNR and BANS for the simulated example under the scenario of spatial correlations, where the data were generated with the sample size $n = 400$, the number of nodes $p = 50$, and the number of time points $T = 60$; the average AUC was produced with the hyper-parameters $(a_1, b_1) = (3, 1)$, $(5, 1)$, and $(10, 1)$; and the average computational time was measured in minutes on a computer with i9-10900K CPU@3.6GHz and 128GB of memory, running in parallel with 20 threads.

		JGGM+MNR			BANS
	Parameter	$a_1 = 3, b_1 = 1$	$a_1 = 5, b_1 = 1$	$a_1 = 10, b_1 = 1$	-
$p = 50$	AUC	0.915(0.004)	0.915(0.004)	0.916(0.004)	-
	Time	7.371	7.798	7.639	$>$ 12 hours

§2.4 Varying Connection Strengths for Different Subjects

In this subsection, we assessed the performance of the proposed method under the scenario of varying connection strengths for different subjects. In this case, we relaxed the assumption stated in the main text by assuming that all subjects share the same effective connections but with different connection strengths. Similar simulation studies were done in Smith et al.[8] Here, we set $n = 400$, $p = 50$, and $T = 60$. A strong signal setting was used for the simulation with $\rho = 0.9$ and the parameters being drawn uniformly from the interval

$[-0.9, -0.6] \bigcup [0.6, 0.9]$. Then, the transition matrix $\boldsymbol{A}_{t,l} \in \mathbb{R}^{p \times p}$ was generated in the procedure given in the section §2.1. For each subject i, we added a random noise $e_t^{(i)} \sim N(0, 0.25)$ to each non-zero entry of $\boldsymbol{A}_{t,l}$ at each point time t. We denoted by $\boldsymbol{A}_{t,l}^{(i)}$ the transition matrix for subject i, for $i = 1, 2, \ldots, n_t$.

The data sets were generated from equations (9) to (11) of the main text with $L = 1$ and $\boldsymbol{\epsilon}_t^{(i)} \sim N_p(0, \Sigma)$, where $\Sigma = \text{diag}\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\}$, and each σ_j^2 was drawn from the uniform distribution $\mathcal{U}(0.9, 1.2)$. The numerical results were summarized in Table S4.

Table S4: Comparison of JGGM+MNR and BANS for the simulated example under the scenario of varying connection strengths for different subjects, where the data were generated with the sample size $n = 400$, the number of nodes $p = 50$, and the number of time points $T = 60$; the average AUC was produced with the hyper-parameters $(a_1, b_1) = (3, 1)$, $(5, 1)$, and $(10, 1)$; and the average computational time was measured in minutes on a computer with i9-10900K CPU@3.6GHz and 128GB of memory, running in parallel with 20 threads.

		JGGM+MNR			BANS
	Parameter	$a_1 = 3, b_1 = 1$	$a_1 = 5, b_1 = 1$	$a_1 = 10, b_1 = 1$	-
$p = 50$	AUC	0.950(0.004)	0.950(0.004)	0.950(0.004)	-
	Time	7.776	7.405	7.478	$>$ 12 hours

§2.5 Computational Complexity

This subsection provides detailed reports on the time complexity regressions discussed in Section 5.1 of the main text. The results reveal that for the BANS method, both the sample size n and the dimension p have a significant impact on the fitted time complexity regression. However, for the JGGM+MNR method, only the dimension p is found to be significant. These findings suggest that the time complexity of the proposed method is not sensitive to the sample size n and demonstrates excellent scalability in relation to the dimension p.

Table S5: Reports for the time complexity regression: for BANS, the regression is given by $\log(\text{Time}) = -1.189 + 0.007n + 0.104p$ with $R^2 = 0.985$; and for JGGM+MNR, the regression is given by $\text{Time} = 3.466 + 0.0003n + 0.078p$ with $R^2 = 0.983$.

Method	parameter	Coefficient	(Standard Deviation)	p-value
BANS	n	0.007	(0.00045)	3.28×10^{-9}
	p	0.104	(0.0044)	2.61×10^{-11}
JGGM+MNR	n	0.0003	(0.00031)	0.376
	p	0.078	(0.003)	6.33×10^{-12}

§2.6 High-Dimensional and Large-Scale Data Case

In this subsection, we conducted simulation studies under the large p and large n scenarios to demonstrate the potential of the proposed method for learning whole brain networks. The data generation procedures and simulation settings are the same as described in Section §2.1. Two cases were considered here: (i) $n = 800$, $p = 300$, and $T = 60$, (ii) $n = 400$, $p = 500$ and $T = 60$. Case (i) represents a large dataset, whose scale is close to a real fMRI dataset, while case (ii) represents a small-n-large-p dataset. In case (ii), to illustrate the robustness of the proposed method, we simulated the transition matrices with varying connection strengths for different subjects (see Section §2.4). For both cases, we simulated data under the strong signal setting.

For both cases, we compared the proposed method with some existing regularization methods, including Lasso,[9] Elastic Net,[10] and MCP,[11] which fit a high-dimensional linear regression separately for each ROI from $t-1$ to t. The numerical results are summarized in Figure S1, which suggests that the proposed JGGM+MNR method outperforms regularization methods significantly in both cases.

§2.7 Network Structure Estimation

In this subsection, we compared the true and estimated Bayesian networks for some time points, see Figure S2 and Figure S3.

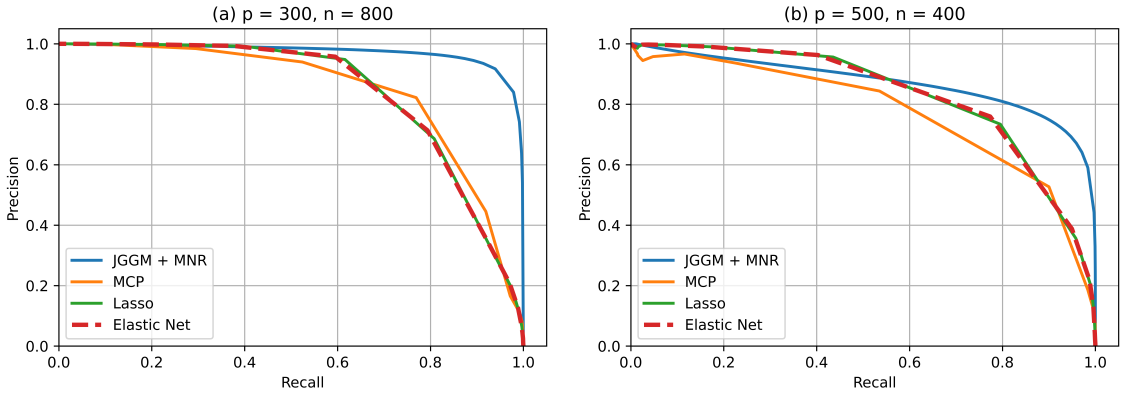

Figure S1: Precision-recall curves produced by JGGM+MNR, MCP, Lasso, and Elastic Net for the simulated examples: (a) results with $p = 300$, $n = 800$ and $T = 60$, where the average AUCs produced by JGGM+MNR, MCP, Lasso, and Elastic Net are 0.972(0.001), 0.844(0.002), 0.842(0.005), and 0.840(0.005), respectively. The average computational time is 1.24 hours for JGGM+MNR and 0.12 hours for Lasso, Elastic Net, and MCP; (b) results with $p = 500$, $n = 400$ and $T = 60$, where the average AUCs produced by JGGM+MNR, MCP, Lasso, and Elastic Net are 0.874(0.001), 0.772(0.002), 0.827(0.004), and 0.827(0.004), respectively. The average computational time is 2.34 hours for JGGM+MNR, 0.42 hours for MCP, and 0.22 hours for Lasso and Elastic Net.

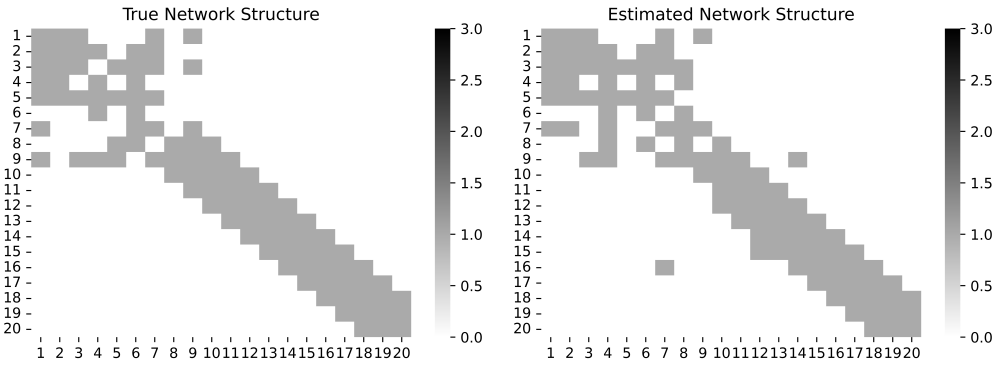

Figure S2: Dynamic Bayesian networks for the transition from $T = 7$ to $T = 8$, where the data were generated with $\rho = 0.9$, $p = 20$, and $n = 400$: (Left) true Bayesian network; (right) estimated Bayesian network.

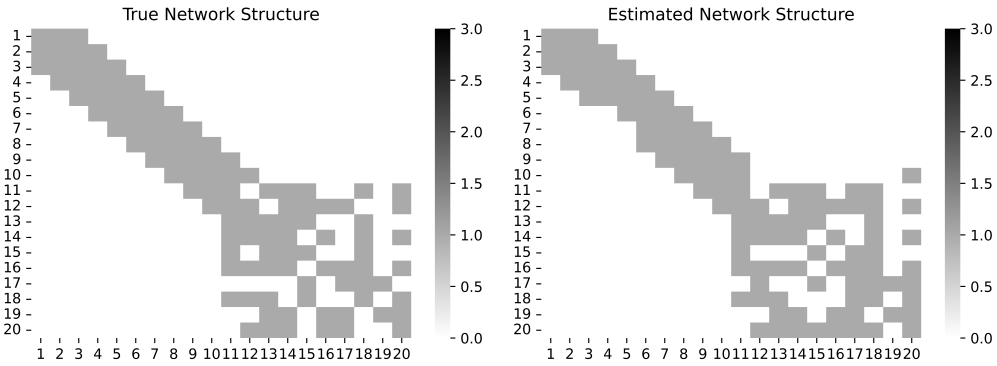

Figure S3: Dynamic Bayesian networks for the transition from $T = 59$ to $T = 60$, where the data were generated with $\rho = 0.9$, $p = 20$, and $n = 400$: (left)true Bayesian network; (right) estimated Bayesian network.

§2.8 A Simulation Study for Lag Order $L = 2$

The data sets were simulated using equations (9) to (11) of the main text with the lag order $L = 2$. We assume that there was one task only. Consequently, all the connections were generated among the regions of interest (ROIs) $1 : p/2$. We set $n = 1000$, $p = 50$, and $T = 61$. The transition matrix $\boldsymbol{A}_{0,l}$ was initialized with $\rho = 0.9$. Then we employed the procedures described in section 5.1 of the main text to generate the time-varying transition matrices $\boldsymbol{A}_{t,l}$'s for $l = 2, 3, \ldots, T$. We added extra 10% non-zero connections in the blocked submatrix $\boldsymbol{A}_{0,l}[1 : \frac{p}{2}, 1 : \frac{p}{2}]$. The parameters were drawn uniformly from the interval $[-0.9, -0.6] \bigcup [0.6, 0.9]$. During the task period for $t = 1, 2, \cdots, 61$, we changed 5% entries of the blocked submatrix $\boldsymbol{A}_{t,l}[1 : \frac{p}{2}, 1 : \frac{p}{2}]$ to zero and another 5% of its entries to non-zero, except for the diagonal elements. The non-zero elements were drawn uniformly from the interval $[-0.9, -0.6] \bigcup [0.6, 0.9]$. The random errors $\boldsymbol{\epsilon}_t^{(i)}$ were drawn from $N_p(0, \Sigma)$, where $\Sigma = \text{diag}\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\}$ and each $\sigma_j^2 \sim \mathcal{U}(0.9, 1.2)$. In particular, when generating the parameter matrix $\boldsymbol{A}_{t,2}$ for lag order $L = 2$, we considered the following 2 cases:

Let $\boldsymbol{B}_{t,1} = \{(i, j) : \boldsymbol{A}_{t,1}{}^{(i,j)} \neq 0, i, j = 1, 2, \ldots, p\}$ and $\boldsymbol{B}_{t,2} = \{(i, j) : \boldsymbol{A}_{t,2}{}^{(i,j)} \neq 0, i, j = 1, 2, \ldots, p\}$ be the set of locations of non-zero elements in the transition matrix $\boldsymbol{A}_{t,1}$ and $\boldsymbol{A}_{t,2}$, respectively. The first case is $\boldsymbol{B}_{t,2} \subset \boldsymbol{B}_{t,1}$ and the second case is $\boldsymbol{B}_{t,1} \cap \boldsymbol{B}_{t,2} = \emptyset$. The numerical results were summarized in Table S6, and the AUC curves were shown in Figure S4. The performance of the proposed method is better in the case $\boldsymbol{B}_{t,1} \cap \boldsymbol{B}_{t,2} = \emptyset$. However,

from the biological side, we do believe that $\boldsymbol{B}_{t,2} \subset \boldsymbol{B}_{t,1}$ is a more reasonable assumption for brain functional connectivity.

Table S6: Performance of JGGM+MNR for the simulated examples with $L = 2$, $n = 1000$, $p = 50$, and $T = 61$. The average AUC was produced with hyper-parameters $a_1 = 10$ and $b_1 = 1$, and the average computational time was measured in minutes on a computer with i9-10900K CPU@3.6GHz and 128GB of memory, running in parallel with 20 threads.

		$\boldsymbol{B}_{t,2} \subset \boldsymbol{B}_{t,1}$	$\boldsymbol{B}_{t,1} \cap \boldsymbol{B}_{t,2} = \emptyset$
$p = 50$	AUC	0.764(0.004)	0.790(0.008)
	Time	16.07	15.41

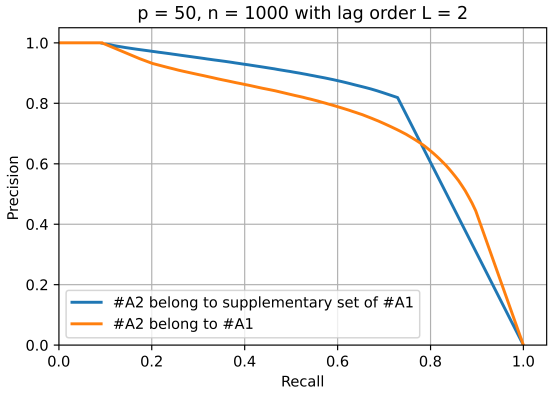

Figure S4: Precision-recall curves produced by JGGM+MNR for simulated examples with $L = 2$, $n = 1000$, $p = 50$, and $T = 61$.

§2.9 A Simulation Study for Less Sparse Transition Matrices

This subsection assesses the effect of sparse levels of the transition matrix on the performance of the proposed method. In Table 1 of the main text, we evaluated the performance of the proposed method with the sparsity level of the transition matrix set to 10%. In what follows, we conducted an experiment with the sparsity level of the transition matrix set to 20%.

In the simulation, we exchanged 10% of the entries of the transition matrix between zero and nonzero values. This exchange corresponds to changes in the true network structure.

Additionally, we set $L = 1$, $n = 400$, $p = 50$, and $T = 60$ under the strong signal case with $\rho = 0.9$. The data was generated as described in section §2.1 of the supplementary material. The numerical results were summarized in Table S7, which indicates that the method performs reasonably well even with more dynamic changes in the network structure at each time point.

Table S7: Average AUCs produced by JGGM+MNR for the simulated examples with $p = 50$, $n = 400$, $T = 60$, and $\rho = 0.9$, under the hyperparameter settings $\alpha_1 = 0.2$ and $(a_1, b_1) = (3, 1)$, (5,1), and (10,1). The average computational time was 8.498 minutes, measured on a computer with i9-10900k CPU@3.6GHz and 128G memory, running in parallel with 20 threads.

Parameter	$a_1 = 3, b_1 = 1$	$a_1 = 5, b_1 = 1$	$a_1 = 10, b_1 = 1$
AUC	0.927(0.002)	0.927(0.003)	0.926(0.003)

§2.10 Sensitivity Analysis for α_1

This subsection assessed the sensitivity of the proposed method with respect to the choice of α_1. We tried $\alpha_1 = 0.05, 0.1, 0.2, 0.25, 0.3$. Additionally, we set the percentage of non-zero entries of the transition matrix to 20%, and exchanged 10% entries between zero and nonzero at each time t; we set $L = 1$, $n = 600$, $p = 50$, and $T = 60$ under the strong signal case with $\rho = 0.9$. The data was generated as described in section §2.1 of the supplementary material.

The numerical results were summarized in Table S8, which suggests that the performance of the proposed method is very robust to the choice of a_1, while it can be affected by the choice of α_1. As analyzed in the main text, a slightly large value of α_1, such as 0.2, is recommended. An excessively large value of α_1 will incur a higher computational cost while providing only marginal improvement in estimation accuracy.

Table S8: Average AUCs produced by JGGM+MNR for the simulated examples with $p = 50$, $n = 600$, $T = 60$, and $\rho = 0.9$, under the hyperparameter settings $\alpha_1 = 0.05, 0.1, 0.2, 0.25, 0.3$ and $(a_1, b_1) = (5, 1)$, $(10, 1)$, and $(15, 1)$. The average computational time was 8.252 minutes, measured on a computer with i9-10900k CPU@3.6GHz and 128G memory, running in parallel with 20 threads.

Parameter	$a_1 = 5, b_1 = 1$	$a_1 = 10, b_1 = 1$	$a_1 = 15, b_1 = 1$
$\alpha_1 = 0.05$	0.944(0.006)	0.943(0.006)	0.943(0.006)
$\alpha_1 = 0.1$	0.948(0.005)	0.948(0.005)	0.947(0.005)
$\alpha_1 = 0.2$	0.954(0.004)	0.954(0.004)	0.953(0.005)
$\alpha_1 = 0.25$	0.956(0.004)	0.956(0.004)	0.956(0.004)
$\alpha_1 = 0.3$	0.958(0.004)	0.957(0.004)	0.957(0.004)

§2.11 Histograms for Z-scores

This subsection presents some histograms of the z-scores calculated for a simulated example with $p = 50$ and $T = 200$. Figure S5 shows the histograms of the z-scores for two selected edges, which indicate that it is reasonable to assume that the z-scores of each edge follow a two-component mixture Gaussian distribution.

Additionally, we plotted the histogram of the z-scores for all edges in Figure S6. It indicates that the z-scores of each edge following a two-component mixture Gaussian distribution is indeed a reasonable assumption. Furthermore, it also indicates that the mean value of the non-zero mean component can be different for different edges. This aligns well with the assumption we made in the paper.

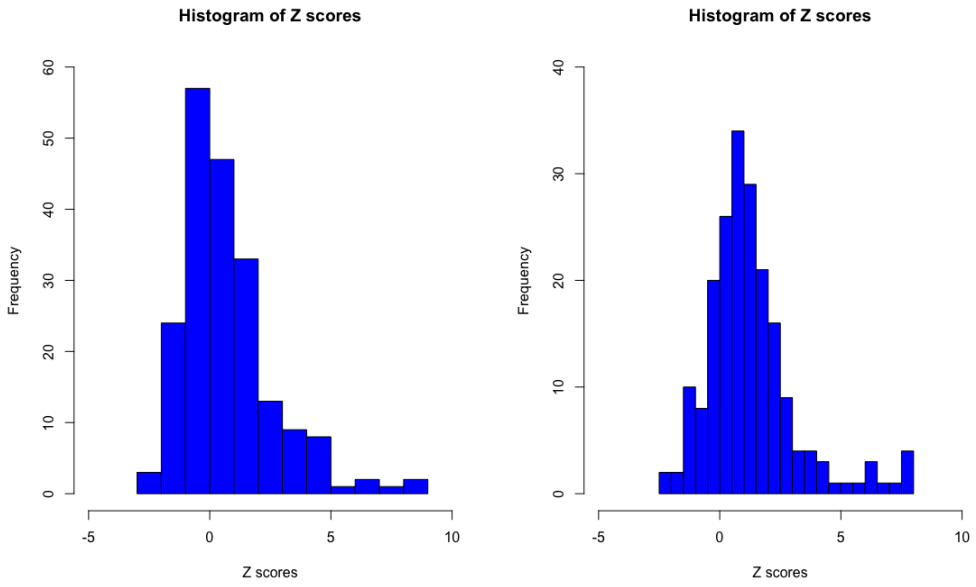

Figure S5: Histograms of the z-scores for two selected edges in a simulated example with $p = 50$ and $T = 200$.

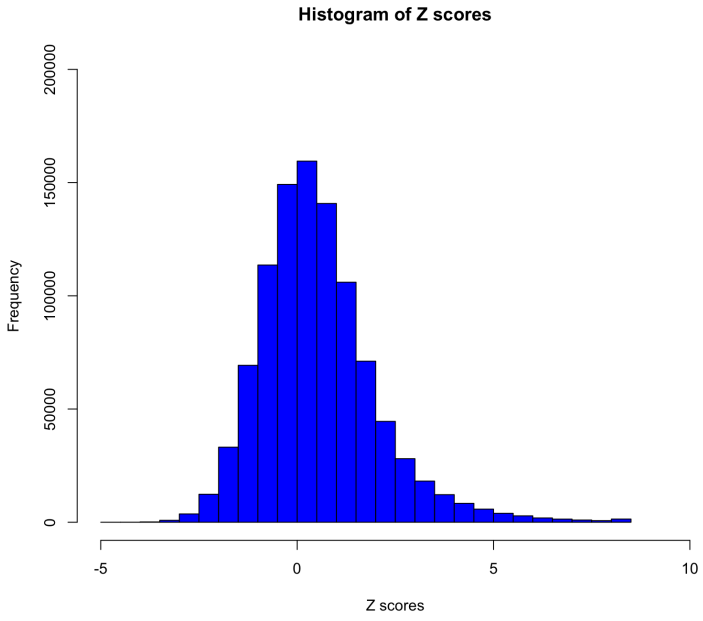

Figure S6: Histogram of the z-scores for all edges in a simulated example with $p = 50$ and $T = 200$.

§3 An fMRI Study of Emotion Processing

This subsection presents the results of the proposed method under different parameter settings, which can be viewed as the sensitivity analysis for the proposed method. The 268 ROIs are divided into 8 functional networks (FN) modules (see Figure S7) including Medial Frontal (Med F), Frontoparietal (FP), Default Mode (DMN), Subcortical-cerebellum (Sub-Cer), Motor (Mt), Visual I (Vis I), Visual II (Vis II) and Visual Association (Vis Assn) as in Cai et al.[12]

(a) Medial Frontal (b) Frontoparietal (c) Default Mode (d) Subcortical-Cerebellum

(e) Motor (f) Visual I (g) Visual II (h) Visual Association

Figure S7: Visualizations of the ROIs in eight functional networks.

§3.1 Results with Different Choices of α_1

Figure S8 shows the heat map of the mean edge (functional module-wise) difference between the emotion and shape tasks with a significance level of $\alpha_1 = 0.05, 0.1, 0.2$ and $\alpha_2 = 0.05$ being used in joint edge detection. Compared to the case with $\alpha_2 = 0.1$ (presented in the main text), fewer connections are detected with $\alpha_2 = 0.05$. However, the conclusion we draw at $\alpha_2 = 0.1$, i.e., the intra-modular connectivity of Sub-Cer and the inter-modular connectivity Sub-Cer to Mt, Sub-Cer to Vis II and Sub-Cer to Vis Assn are different under

the emotion and shape tasks, still holds. Additionally, the conclusion holds for different choices of α_1.

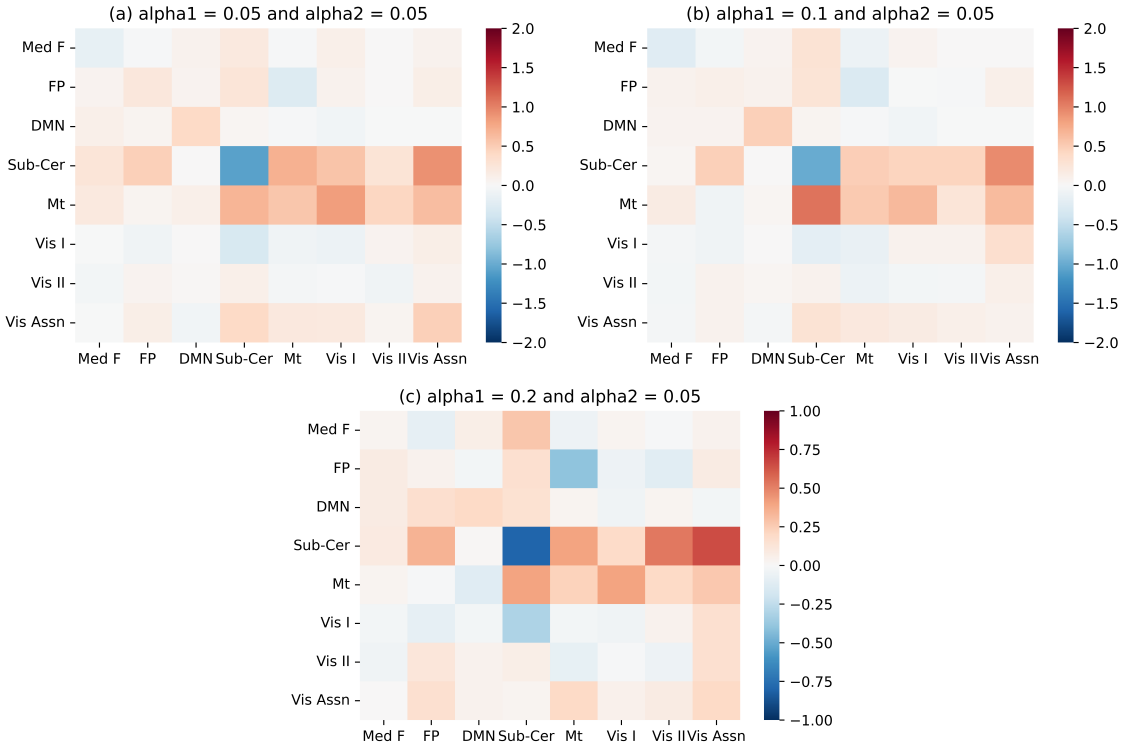

Figure S8: Heat maps of the mean edge (functional module-wise) difference between the emotion and shape tasks, which were identified with (a) $(\alpha_1, \alpha_2) = (0.05, 0.05)$, (b) $(\alpha_1, \alpha_2) = (0.1, 0.05)$, and (c) $(\alpha_1, \alpha_2) = (0.05, 0.2)$.

§3.1.1 Identification of Emotion-related ROIs

The ROI was identified using the Wilcoxon signed-rank test with a significance level of $\alpha = 0.05$. For the learned dynamic Bayesian networks, we summarized the edges stemming from each ROI at each time t and under each task. Therefore, for each ROI, we obtained a sequence of edge numbers under each of the emotion and shape tasks. For each ROI, we tested the median difference of the two sequences using the Wilcoxon signed-rank test. Then, we adjusted the p-values by a multiple-hypothesis test using the empirical Bayesian method of Liang and Zhang (2008),[13] which is available in the R package *equSA*[14]. The ROI 32 shows a small Storey's q-value[15] of $q = 0.000131$, which indicates that this ROI performs

differentially under the emotion and shape tasks. Notably, ROI 32 was the only ROI with a Storey's q-value less than 0.05 in our test. The location information about ROI 32 is given in Table S9 and Figure S9 displays the variation of row degrees along with the time, which tends to have higher connectivity under the emotion task than under the shape task.

Finally, we note that the Wilcoxon rank-sum test can be used in the test if we ignore the time correspondence between the two tasks.

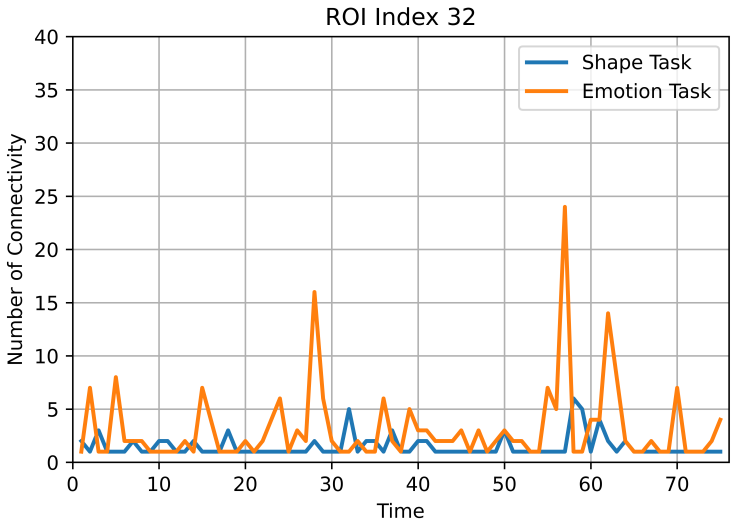

Figure S9: Numbers of coming-out-edges from ROI 32 in the time-varying dynamic Bayesian networks learned by JGGM+MNR for different tasks.

Table S9: Anatomical location, functional network module, and MNI coordinates of the identified hub ROIs.

			MNI			
Index	Lobes	Region	X	Y	Z	Network
32	R-MotorStrip	BA6	32	-5.4	52.1	Sub-Cer

§3.2 Results with Poisson HRF

Other than the Canonical HRF function, we also tried the Poisson HRF[16] by setting $h_{\lambda_v} = \exp(-\lambda_v)\lambda_v^t/t!$, where the parameter λ_v is drawn from $\mathcal{U}(0, 8)$. The results are summarized in Figure S10, which are similar to those obtained with canonical HRF.

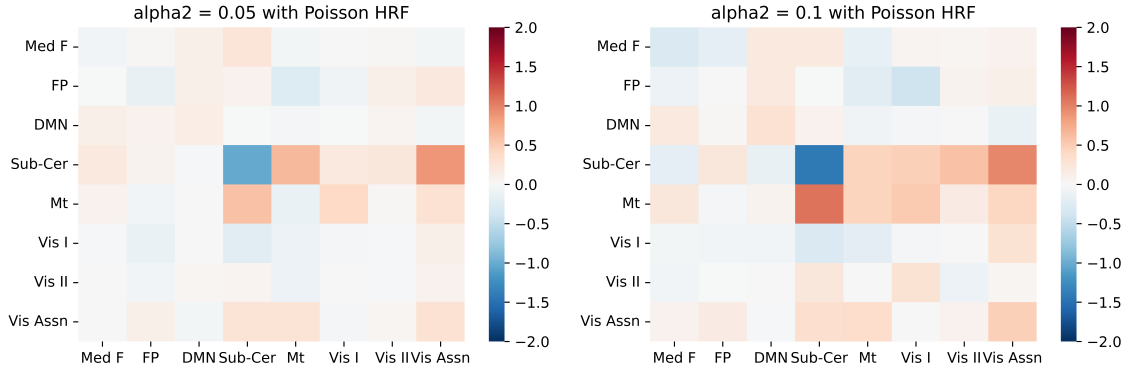

Figure S10: Heat maps of the mean edge difference (functional module-wise) between emotion and shape tasks with Poisson HRF ($\lambda_v = 0.739$), where the left plot was obtained with $\alpha_2 = 0.05$, and the right plot was obtained with $\alpha_2 = 0.1$.

§3.3 Results with the Lag Order $L = 2$

Following the data analysis pipeline as described in Section 6.3, we re-analyzed the data with the lag order $L = 2$. Figure S11 shows the resulting heat maps. Our results indicate that less connectivity is detected with $L = 2$, which suggests the high-order Markov nature of brain activity.

However, even with $L = 2$, we found that the difference in the intra-connectivity of the subcortical cerebellum can still be detected. This suggests consistent temporal differences between shape and emotion tasks.

§3.4 Results by Regularized Methods

For the regularization methods, the regularization parameter λ was chosen to ensure the number of connections selected at each time point is similar to those selected by the proposed method. The regularized methods were implemented with the R package **ncvreg**,[17] and the

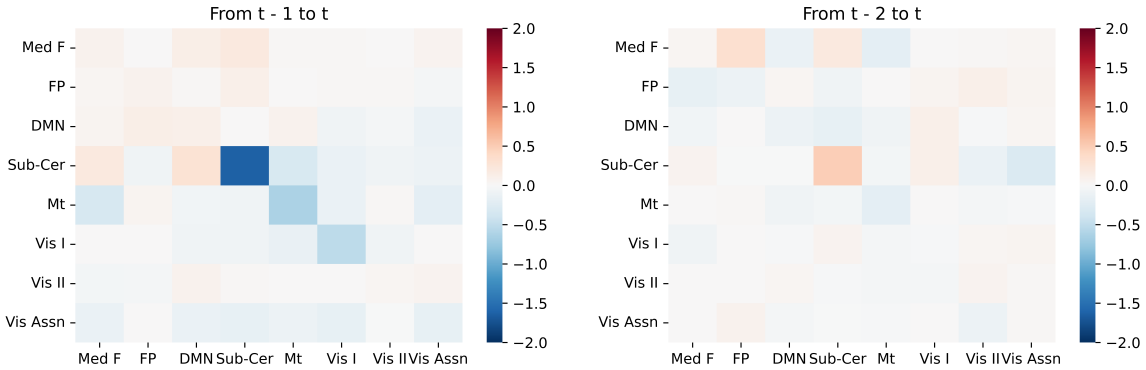

Figure S11: Heat maps for the mean edge difference (functional module-wise) under the emotion and shape tasks, which were obtained with $L = 2$: (left) heat map for the difference of connectivity from $t - 1$ to t, and (right) heat map for the difference of connectivity from $t - 2$ to t.

results were summarized in Figure S12, which suggests that the connectivity between the emotion and shape tasks is not much different, except for the intra-connectivity in the Vis Assn module.

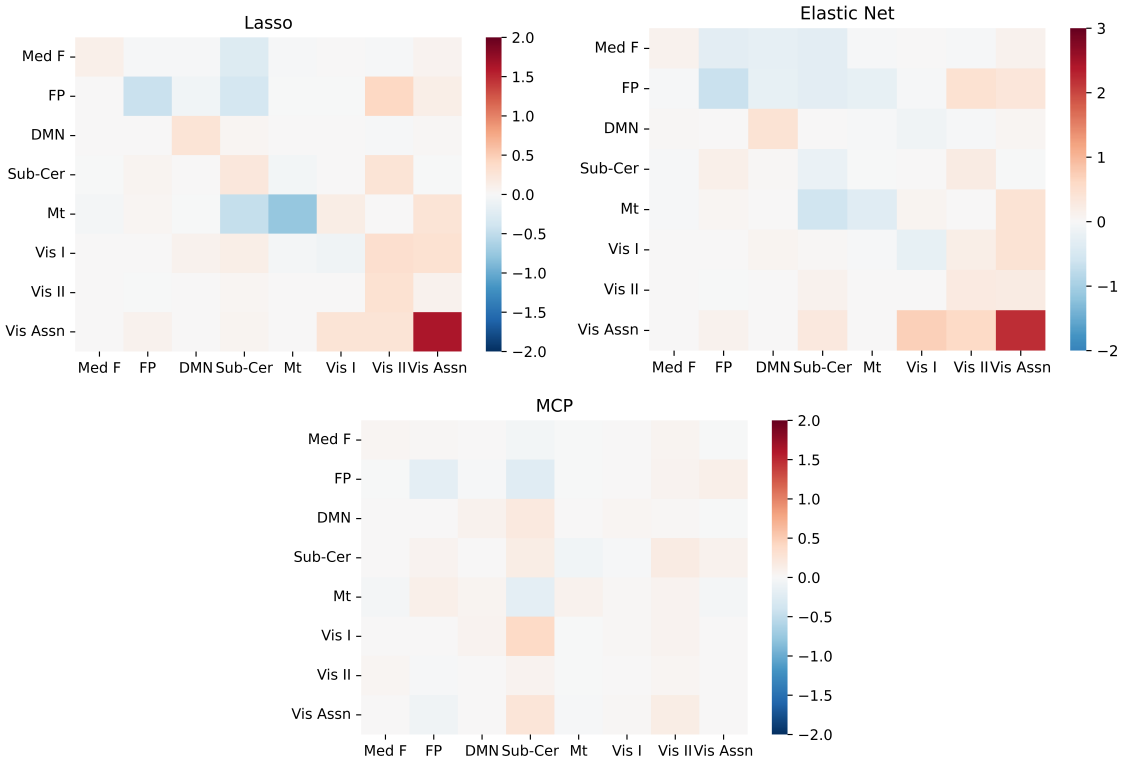

Figure S12: Heat maps for the mean edge difference (functional module-wise) between emotion and shape tasks by the regularized methods: Lasso, Elastic Net, and MCP.

References

[1] Liang F., Xue J., Jia B.. Markov Neighborhood Regression for High-Dimensional Inference *Journal of the American Statistical Association.* 2022;117:1200-1214.

[2] Jia Bochao, Liang Faming, TEDDY Study Group . Fast hybrid Bayesian integrative learning of multiple gene regulatory networks for type 1 diabetes *Biostatistics.* 2021;22:233-249.

[3] Liang F., Song Q., Qiu P.. An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models *Journal of the American Statistical Association.* 2015;110:1248-1265.

[4] Fan Jianqing, Peng Heng. Nonconcave penalized likelihood with a diverging number of parameters *Ann. Statist..* 2004;32:928–961.

[5] Fan J., Guo S., Hao N.. Variance estimation using refitted cross-validation in ultrahigh dimensional regression *Journal of the Royal Statistical Society, Series B.* 2012;74:37-65.

[6] Fan Jianqing, Lv Jinchi. Sure independence screening for ultrahigh dimensional feature space *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2008;70:849–911.

[7] Shen Xilin, Tokoglu Fuyuze, Papademetris Xenios, Constable R Todd. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification *Neuroimage.* 2013;82:403–415.

[8] Smith Stephen M., Miller Karla L., Khorshidi Gholamreza Salimi, et al. Network modelling methods for FMRI *NeuroImage.* 2011;54:875-891.

[9] Tibshirani Robert. Regression shrinkage and selection via the lasso *Journal of the Royal Statistical Society. Series B (Methodological).* 1996:267–288.

[10] Zou Hui, Hastie Trevor. Regularization and variable selection via the elastic net *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2005;67:301–320.

[11] Zhang Cun-Hui. Nearly unbiased variable selection under minimax concave penalty *Annals of Statistics.* 2010:894–942.

[12] Cai Biao, Zhang Gemeng, Zhang Aiying, et al. Functional connectome fingerprinting: identifying individuals and predicting cognitive functions via autoencoder *Human Brain Mapping.* 2021;42:2691–2705.

[13] Liang Faming, Zhang Jian. Estimating the false discovery rate using the stochastic approximation algorithm *Biometrika.* 2008;95:961–977.

[14] Jia Bochao, Liang Faming, Shi Runmin, Xu Suwa. equSA: Estimate Directed and Undirected Graphical Models and Construct Networks *CRAN-Package.* 2018.

[15] Storey J.D.. A direct approach to false discovery rates *Journal of the Royal Statistical Society, Series B.* 2002;64:479-498.

[16] Warnick Ryan, Guindani M., Erhardt Erik B., Allen Elena A., Calhoun Vince D., Vannucci Marina. A Bayesian Approach for Estimating Dynamic Functional Network Connectivity in fMRI Data *Journal of the American Statistical Association.* 2018;113:134 - 151.

[17] Breheny Patrick, Huang Jian. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection *The annals of applied statistics.* 2011;5:232.