Received: 11 April 2021 | Revised: 25 June 2021 | Accepted: 18 August 2021
DOI: 10.1002/sta4.414

SPECIAL ISSUE PAPER

WILEY

Structural inference of time-varying mixed graphical models

Qingyang Liu[1] | Yuping Zhang[1] | Zhengqing Ouyang[2]

[1]Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

[2]Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, Amherst, MA 01003, USA

Correspondence
Yuping Zhang, Department of Statistics, University of Connecticut, Storrs, CT 06269, USA.
Email: yuping.zhang@uconn.edu

Funding information
Division of Mathematical Sciences, Grant/Award Number: 2015481

With the rapid advancement of biotechnology, there arise new challenges to discover regulatory and co-action relationships among heterogeneous biological features as well as to capture significant topological changes when data are observed across time. This paper is devoted to joint structural estimation of time-varying mixed graphical models based on multivariate data over a series of time points. Assuming the graph topology changes gradually over time, we establish a flexible local estimator to fully exploit the structural smoothness. Utilizing variational likelihood inference, we impose a group lasso penalty to integrate information from nearby time points. In order to reduce the algorithmic complexity, we propose an accelerated alternating direction method of multipliers (ADMM)-based algorithm exploiting the block diagonal structure to adapt our problem for large sparse networks. Practical merits of our method are exhibited through synthetic networks with mixed data types. Ultimately, we illustrate the real application by incorporating multi-platform data from the PsychENCODE human brain development project and detect the evolution of gene regulatory networks along different stages of human brain development.

KEYWORDS
ADMM, brain development, DNA methylation, group lasso, joint modelling, network

1 | INTRODUCTION

In biomedical research, reconstruction of correlation networks between different types of features is a fundamental step to understand the functions and mechanisms among them. As for multi-omic studies, a variety of genomic, transcriptomic, proteomic or epigenomic data are integrated together to form a comprehensive view of the studied organisms at the molecular level. Common biological networks include protein–protein interaction (PPI) networks, gene co-expression networks and gene regulatory networks, classified by the types of molecules involved. Gene regulatory networks study the interactions (i.e., activations and inhibitions) between transcription factors and expressions of genes, which play an important role in numerous cellular processes (Karlebach & Shamir, 2008). PPI networks (Chautard et al., 2009) as well as gene co-expression networks (Zhang & Horvath, 2005) catalogue mutual interactions among proteins and gene expressions, respectively. The inference of biological networks is critical in detecting important co-action and causal relationships among different molecules, so as to discover gene or protein functions, signalling pathways and causation of many diseases. With the rapid advancement of biotechnology, there emerge new experimental and observational studies to reveal the mechanisms of many progression processes, such as cell differentiation, cell cycle and brain development, at the cell or tissue level. To better uncover the evolution of molecule interactions in complex biological systems, it is of great necessity to model a sequence of temporal dynamic networks and portray the evolution trajectory.

Undirected graphical models, also known as Markov Random Fields, is a widely used model-based approach to characterize the conditional dependency structure for multivariate data, including representative examples such as Gaussian graphical models for symmetric and thin-tailed continuous data, as well as Ising models for binary data. To extend the application of graphical models to data composed of heterogeneous types of features, Lauritzen and Wermuth (1989) proposed an undirected graphical model for mixtures of categorical and Gaussian variables. Then,

https://doi.org/10.1002/sta4.414

Cheng et al. (2017) and Lee and Hastie (2015) simplified the model to improve the scalability for a large amount of categorical variables. In addition, a line of work developed a sub-class of Markov Random Fields (Chen et al., 2014; Tansey et al., 2015; Yang et al., 2012, 2014), which is specified by conditional distributions belonging to potentially different exponential families. Such mixed graphical models are useful in the vertical integration of multi-platform data, for example, gene expressions, mutations, copy number variations and epigenetic states, including binary, categorical, count and continuous variables. In terms of horizontal integration of multiple static conditions, Danaher et al. (2014) developed the joint graphical lasso to estimate multiple Gaussian graphical models. Furthermore, a line of work developed multiple mixed graphical models (Liu & Zhang, 2020b; Zhang et al., 2017) for simultaneously vertical integration of heterogenous data types and horizontal integration of multiple static conditions.

This paper is aimed at the situation that the same set of heterogeneous features are observed over a temporal grid, and we would like to estimate a series of temporal dynamic graphical models to elucidate the evolution process. For instance, modelling time-varying regulation networks of human brain development can help targeting critical transcription factors and gene markers at different stages of brain development. For multivariate data observed on temporal grids, abundant existing works (Gibberd & Nelson, 2017; Gibberd & Roy, 2017; Hallac et al., 2017; Monti et al., 2014; Qiu et al., 2016; Yang & Peng, 2020; Zhou et al., 2010) proposed different frameworks to jointly estimate smoothly evolving Gaussian graphical models; Song et al. (2009) and Kolar et al. (2010) developed methodologies for time-varying graphical models using binary and discrete data. To impose temporal smoothness, we have to make assumptions on how the true models vary over time. One type of local stationarity is the model parameters are smooth functions of time, which can be implemented by kernel smoothing to combine observations close in time. The other type of assumption is the model parameters are piecewise constant, often achieved by fused lasso type penalties. Recently, Haslbeck and Waldorp (2020) developed an R package named **mgm** to estimate pairwise and higher order interactions in mixed graphical models, which also supports time-varying mixed graphical models. It is capable of handling mixtures of three types of variables: Gaussian, count and categorical. The temporal smoothness is introduced by a Gaussian kernel on nodewise regressions. The nodewise regressions employed in **mgm** yield asymmetric estimates of edge parameters, and the final topology estimates are formed by "AND" or "OR" rule to unify all neighbourhood structures.

In this paper, we propose the "*Local Approximate Likelihood Data Integration*" (LoALDIG), a joint structural inference framework for temporal dynamic mixed graphical models, fully taking advantage of the smoothness of dependency structures. Besides assuming smoothness in model parameters, we make another reasonable assumption that the graph topology gradually changes over time, which is very beneficial in recovering edge structures. Thus, in additional to kernel weighting on approximate likelihood, we introduce a local group lasso regularization with adjustable width. In comparison with **mgm**, not only does our approach borrow information from neighbouring time points, but it also adapts to the local degree of smoothness in a data-driven manner. Furthermore, our LoALDIG avoids the common asymmetry problem of nodewise regression type of approaches.

To organize the rest of the paper, we start from elaborating the statistical model as well as the joint modelling framework in Section 2. Then, we propose an algorithm for model fitting in Section 3 and introduce a purely data-driven strategy for tuning parameter selection in Section 4. Practising our method via numerical experiments in Section 5, we demonstrate the practical merits of LoALDIG. A case study on the human brain development data is carried out in Section 6, followed by conclusions in Section 7.

2 | METHODS

Suppose $\boldsymbol{x}(t_k) = (x_1(t_k),...,x_p(t_k))^\top \in \mathbb{R}^p$ denotes a p-variate random vector observed at time point t_k, following a joint distribution $f(\boldsymbol{x};\boldsymbol{\theta}(t_k))$. In practice, we rescale all time points, so that $t_k \in [0,1]$ for $k = 1,...,K$. The set of p features shared across the timeline are potentially heterogeneous, possessing different supports and measures from each other, such as categorical, Gaussian, Poisson, truncated Poisson (Yang et al., 2013) and exponential. In this paper, we consider not only time series data but also data with repeated measurements at each observed time point. Therefore, we denote the full data matrix at t_k by $\boldsymbol{X}(t_k) = (\boldsymbol{x}^1(t_k),...,\boldsymbol{x}^{n_k}(t_k))$, and the n_k observations are assumed independent and identically distributed. In addition, we assume that $\{\boldsymbol{X}(t_1),...,\boldsymbol{X}(t_K)\}$ are independent along time, but the underlying model parameters $\boldsymbol{\theta}(t_k)$'s have a smooth trend of change temporally.

2.1 | Pairwise exponential Markov Random Field and its variational likelihood inference

Omitting all indices indicating time for simplicity in this part, we first define the joint distribution $f(\boldsymbol{x};\boldsymbol{\theta})$ in the form of pairwise exponential Markov Random Field (Park et al., 2017):

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left\{ \sum_{r=1}^{p} \boldsymbol{\theta}_r^\top \boldsymbol{B}_r(x_r) + \sum_{r=1}^{p}\sum_{s=1}^{p} \langle \boldsymbol{\theta}_{rs}, \boldsymbol{B}_r(x_r)\boldsymbol{B}_s(x_s)^\top \rangle_F + \sum_{r=1}^{p} C_r(x_r) - A(\boldsymbol{\theta}) \right\}, \tag{1}$$

where $\langle \boldsymbol{A},\boldsymbol{B}\rangle_F = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ is the Frobenius inner product, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_p,\boldsymbol{\theta}_{11},\ldots,\boldsymbol{\theta}_{pp}\}$ is the set of natural parameters and $\boldsymbol{B}(\boldsymbol{x}) = \left\{\{\boldsymbol{B}_r(x_r)\}_{r=1}^p, \{\boldsymbol{B}_r(x_r)\boldsymbol{B}_s(x_s)^\top\}_{r,s=1}^p\right\}$ is the set of sufficient statistics. Among the p heterogeneous features, the dimensions of these sufficient statistics and natural parameters may not be the same, so we set $\boldsymbol{\theta}_r \in \mathbb{R}^{m_r}, \boldsymbol{\theta}_{rs} \in \mathbb{R}^{m_r \times m_s}, \boldsymbol{B}_r(x_r) \in \mathbb{R}^{m_r}$. Without loss of generality, we impose symmetry constraints $\boldsymbol{\theta}_{rs} = \boldsymbol{\theta}_{sr}^\top$. This model explicitly specifies a pairwise Markov Random Field factorized by vertices and edges (V, E), including Gaussian graphical models as a special case. That means edge potential $\boldsymbol{\theta}_{rs}$ is a zero matrix if and only if x_r and x_s are conditionally independent given all other variables. $A(\boldsymbol{\theta})$ is the log-partition function which should be finite valued. We are able to verify that the conditional distribution of x_r given all other variables $\boldsymbol{x}_{\setminus r}$ is proportional to

$$\exp\left\{\left(\boldsymbol{\theta}_r + 2\sum_{s \neq r}\boldsymbol{\theta}_{rs}\boldsymbol{B}_s(x_s)\right)^\top \boldsymbol{B}_r(x_r) + \langle \boldsymbol{\theta}_{rr}, \boldsymbol{B}_r(x_r)\boldsymbol{B}_r(x_r)^\top\rangle_F + C_r(x_r)\right\},$$

which is a distribution from the exponential family. Hence, we specify mixed graphical models through full conditional distributions, of which each node can be of a different type (e.g., categorical, Gaussian, exponential, Poisson etc.).

Consider n independent observations, and the log-likelihood can be expressed as $\ell(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}\rangle - A(\boldsymbol{\theta})$, where

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{B}(\boldsymbol{x}^i) = \left\{\left\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_r(x_r^i)\right\}_{r=1}^p, \left\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_r(x_r^i)\boldsymbol{B}_s(x_s^i)^\top\right\}_{r,s=1}^p\right\}$$

is the sample mean sufficient statistics. It is challenging to make direct likelihood inference due to the existence of the intractable log-partition function $A(\boldsymbol{\theta}) = \log\int_{\mathcal{X}}\exp\{\langle \boldsymbol{\theta}, \boldsymbol{B}(\boldsymbol{x})\rangle + C(\boldsymbol{x})\}\mu(\mathrm{d}\boldsymbol{x})$. In general, the log-partition function does not have an analytic form. Consequently, Park et al. (2017) derived a variational method, which utilizes an upper bound of $A(\boldsymbol{\theta})$ to establish an approximate maximum likelihood problem based on the log-determinant relaxation proposed by Jordan and Wainwright (2004) and Wainwright and Jordan (2006).

Because of the Fenchel duality among the maximum likelihood and the maximum entropy (Wainwright & Jordan, 2007), we apply the Gaussian entropy bound to construct the upper bound of $A(\boldsymbol{\theta})$ by introducing the following discreteness adjustment. For each discrete node $r \in \{1, \ldots, p\}$, we denote the minimum distance of sufficient statistic by $c_r = \inf_{a \neq b \in \mathcal{X}_r}\|\boldsymbol{B}_r(a) - \boldsymbol{B}_r(b)\|_\infty > 0$ ($c_r = 0$ for continuous nodes). So as to obtain differential entropy, an additive diagonal matrix $\boldsymbol{D} = \mathrm{diag}([0, \boldsymbol{l}_1, \ldots, \boldsymbol{l}_p]) \in \mathbb{R}^{(q+1)\times(q+1)}$ is defined, where $\boldsymbol{l}_r = (c_r^2/12)\cdot \mathbf{1}_{m_r}$ and $q = \sum_{r=1}^p m_r$. For notational simplicity, we further introduce the following map and alternative parameterization. Given a scalar $\nu \in \mathbb{R}$, a map $\mathbf{M}_\nu[\cdot]$ is formulated as: for $\boldsymbol{\mu} = \left\{\{\boldsymbol{\mu}_r\}_{r=1}^p, \{\boldsymbol{\mu}_{st}\}_{s,t=1}^p\right\} \in \mathbb{R}^{m_1} \times \ldots \mathbb{R}^{m_p} \times \mathbb{R}^{m_1 m_1} \times \ldots \mathbb{R}^{m_p m_p}$,

$$\mathbf{M}_\nu[\boldsymbol{\mu}] = \begin{bmatrix} \nu & \boldsymbol{\mu}_1^\top & \boldsymbol{\mu}_2^\top & \cdots & \boldsymbol{\mu}_p^\top \\ \boldsymbol{\mu}_1 & \boldsymbol{\mu}_{11} & \boldsymbol{\mu}_{12} & \cdots & \boldsymbol{\mu}_{1p} \\ \boldsymbol{\mu}_2 & \boldsymbol{\mu}_{21} & \boldsymbol{\mu}_{22} & & \boldsymbol{\mu}_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\mu}_p & \boldsymbol{\mu}_{p1} & \boldsymbol{\mu}_{p2} & \cdots & \boldsymbol{\mu}_{pp} \end{bmatrix} = \begin{bmatrix} \nu & \boldsymbol{\mu}_\bullet^\top \\ \boldsymbol{\mu}_\bullet & \boldsymbol{\mu}_{\bullet\bullet} \end{bmatrix} \in \mathbb{R}^{(q+1)\times(q+1)};$$

$\boldsymbol{\theta}' = \{\boldsymbol{\theta}_1/2, \ldots, \boldsymbol{\theta}_p/2, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{pp}\}$. Further taking relaxation of the Gaussian entropy bound, we are able to derive

$$A(\boldsymbol{\theta}) \leq \frac{1}{2}\min_{\nu \in \mathbb{R}}\left\{-\mathrm{logdet}\left(-2\mathbf{M}_{1+\nu/2}[\boldsymbol{\theta}']\right) - \nu\right\} - \langle \mathbf{D}, \mathbf{M}_1[\boldsymbol{\theta}']\rangle_F + f_1,$$

where f_1 is a constant only determined by node types. Replacing the log-partition function $A(\boldsymbol{\theta})$ by its upper bound, we attain the following approximated maximum likelihood problem:

$$\min_{\boldsymbol{\theta}}\ -\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\theta}\rangle + A(\boldsymbol{\theta}) \approx \frac{1}{2}\min_{\boldsymbol{\Theta} \succ 0}\left\{\langle \boldsymbol{\Theta}, \mathbf{M}_1[\hat{\boldsymbol{\mu}}] + \boldsymbol{D}\rangle_F - \mathrm{logdet}\boldsymbol{\Theta}\right\} + f_2,$$

where $\boldsymbol{\Theta} = -2\mathbf{M}_{1+\nu/2}[\boldsymbol{\theta}']$ and f_2 is another constant. For convenience, we let $\hat{\boldsymbol{\Sigma}} = \mathbf{M}_1[\hat{\boldsymbol{\mu}}] + \boldsymbol{D}$, and thereby, the approximate likelihood of a static model can be presented as

$$L(\boldsymbol{\theta}; \boldsymbol{X}) = \langle \boldsymbol{\Theta}, \hat{\boldsymbol{\Sigma}}\rangle_F - \mathrm{logdet}\boldsymbol{\Theta}.$$

Choosing the group lasso (Yuan & Lin, 2006) as the penalty function to impose group sparsity on edge potentials, Park et al. (2017) showed that the corresponding regularized approximate likelihood method is able to recover the structure of pairwise exponential graphical models with high accuracy.

2.2 | Local estimator for temporal dynamic mixed graphical models

Data $\{\boldsymbol{X}(t_1),...,\boldsymbol{X}(t_K)\}$ are measured over a temporal grid. At observed time point t_k, the underlying probability mass/density $f(\boldsymbol{x};\boldsymbol{\theta}(t_k))$ is specified by the aforementioned pairwise exponential Markov Random Field (1). With limited data, one cannot estimate a model from a single time point. Hence, assuming that $\boldsymbol{\theta}(t)$ have a continuous and smooth change over time, we employ kernel smoothing to obtain the following stabilized approximate likelihood for the model at t_k,

$$\begin{aligned}\tilde{L}_h(\boldsymbol{\Theta}(t_k)) &= \frac{1}{\sum_{j=1}^{K} n_j \mathcal{K}_h(t_j - t_k)} \sum_{j=1}^{K} n_j \mathcal{K}_h(t_j - t_k) L(\boldsymbol{\theta}(t_k);\boldsymbol{X}(t_j)) \\ &= \langle \boldsymbol{\Theta}(t_k), \tilde{\boldsymbol{\Sigma}}(t_k,h) \rangle_F - \operatorname{logdet}(\boldsymbol{\Theta}(t_k)),\end{aligned}$$

where $\mathcal{K}_h$ is a symmetric non-negative kernel function and $h > 0$ is the bandwidth. The choice of bandwidth involves a trade-off between the sensitivity to time-specific information and the stability of the estimate. As we can see, the smoothed approximate likelihood can be considered as applying kernel smoothing on $\hat{\boldsymbol{\mu}}$, that is, $\tilde{\boldsymbol{\Sigma}}(t_k,h) = \boldsymbol{M}_1[\tilde{\boldsymbol{\mu}}(t_k,h)] + \boldsymbol{D}$, where

$$\tilde{\boldsymbol{\mu}}(t_k,h) = \frac{1}{\sum_{j=1}^{K} n_j \mathcal{K}_h(t_j - t_k)} \sum_{j=1}^{K} n_j \mathcal{K}_h(t_j - t_k) \hat{\boldsymbol{\mu}}(t_j).$$

Please note that the kernel weighted approximate likelihood is also applicable to unobserved time points based on the local stationary assumption.

Our study objective is to reconstruct the graph topology $E(t)$ at a given time point $t \in [0, 1]$ based on the observed data. To utilize the data more efficiently, we assume that the edge set changes gradually over time. Based on this assumption, we further apply a group lasso penalty to conduct joint inference of the sparsity pattern of edge potentials,

$$\begin{aligned}\{\hat{\boldsymbol{\Theta}}(t_k;h,d,\lambda)\}_{k \in \mathcal{N}_d(t)} = \underset{\boldsymbol{\Theta}(t_k) \succ 0, k \in \mathcal{N}_d(t)}{\arg\min} & \frac{1}{\sqrt{|\mathcal{N}_d(t)|}} \sum_{k \in \mathcal{N}_d(t)} \tilde{L}_h(\boldsymbol{\Theta}(t_k)) \\ & + \lambda \sum_{r \neq s} w_{rs} \left(\sum_{k \in \mathcal{N}_d(t)} \|\boldsymbol{\Theta}_{rs}(t_k)\|_F^2 \right)^{1/2},\end{aligned} \tag{2}$$

where $\mathcal{N}_d(t) = \{v \in \{1,...,K\} : |t_v - t| \leq d\}$ represents the indices of time points centred around time t with a half neighbourhood width d; $|\mathcal{N}_d(t)|$ is the number of time points included by the neighbourhood; and $\|\boldsymbol{A}\|_F = [\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A})]^{1/2}$ denotes the Frobenius norm. The transformed model parameter matrices $\boldsymbol{\Theta}(t_k)$'s are also partitioned according to $\boldsymbol{M}_\nu[\cdot]$, and $\boldsymbol{\Theta}_{rs}(t_k)$ indicates the (r,s)th block of the mapping. It is noteworthy that non-edge potentials, including the first row, the first column as well as the diagonal of $\boldsymbol{\Theta}(t_k)$, do not get regularized. The weights w_{rs} are designated as $w_{rs} = \sqrt{m_r \times m_s}$ to balance the penalty for edge-potential matrices of different sizes. The group lasso regularizer integrates the weighted likelihood functions within the neighbourhood and synchronizes all sparsity patterns across time. As a result, the level of smoothness is controlled by the tuning parameter d. The wider neighbourhood we select, the more data from remote time points we integrate, so that the estimated topology would change more smoothly provided that h holds fixed. The factor $\frac{1}{\sqrt{|\mathcal{N}_d(t)|}}$ is introduced to ensure that the problem (2) is adaptive to different d. So we are able to apply the same sequence of λ to a range of different neighbourhood sizes for sparsity control.

In summary, the proposed local estimator (LoALDIG) contains two parts of smoothness, including kernel smoothing of variational likelihood and a joint estimation via structured-sparsity-inducing penalty. Its form is connected with the previous research works by Danaher et al. (2014), Liu and Zhang (2020b) and Zhang et al. (2017), which employ two parts of group lasso penalties in a hierarchical structure for joint estimation of multiple graphical models. Their methods are devoted to the scenario that observed data are from K unordered classes, so they accommodate discrepancy in the estimated edge sets to adjust the level of joint modelling. Differently, given that time labels are known in our problem, our estimator with only one group lasso term forces all $|\mathcal{N}_d(t)|$ edge sets to be identical, while the level of joint modelling is controlled by the neighbourhood width. Furthermore, LoALDIG can be viewed as a generalization of Yang and Peng (2020), which relies on the same set of assumptions to make inference on time-varying Gaussian graphical models. Our contribution is to broaden the application of the local group lasso estimation to data with mixed types of attributes.

Based on the solution to problem (2), it is sufficient to extract the common sparsity pattern, so the estimated graph topology $\hat{E}(t;h,d,\lambda)$ is subsequently determined. Regarding parameter estimation, one problem is that the sparsity-inducing penalty is likely to over-shrink the edge potentials. Hence, for more accurate estimation of model parameters, we further carry out the following model refitting procedure by maximizing the weighted approximate likelihood at time t subject to the constraint of the estimated graph topology:

$$\hat{\boldsymbol{\Theta}}^{\mathrm{rf}}(t;h,d,\lambda) = \underset{\boldsymbol{\Theta}(t)\succ 0}{\arg\min}\ \tilde{L}_h(\boldsymbol{\Theta}(t)), \text{subject to } \boldsymbol{\Theta}_{rs}(t)=\mathbf{0} \text{ for all } (r,s)\notin \hat{E}(t;h,d,\lambda).$$

For any time point between 0 and 1, we are able to first make inference on the graph topology and then refit the model to eliminate overshrinkage supposing there is no rank deficiency.

2.3 | Problem reformulation

As for Gaussian graphical models, people usually do data centralization first to focus on the inference of the precision matrix. Likewise, we reformulate the previous joint inference problem (2) to obtain a simplified equivalent form excluding all node potentials. To begin with, we define the following $q \times q$ modified empirical covariance matrix of all sufficient statistics at time t_k smoothed by the kernel function $\mathcal{K}_h$:

$$\tilde{\boldsymbol{H}}(t_k,h) = [\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet\bullet} - [\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet}[\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet}^{\top} + \mathrm{diag}([\boldsymbol{I}_1,\ldots,\boldsymbol{I}_p]).$$

Based on joint modelling within the local neighbourhood, we consider the following alternative optimization problem for local edge set recovery without inferring node potentials:

$$\begin{aligned}\{\hat{\boldsymbol{\Omega}}(t_k;h,d,\lambda)\}_{k\in\mathcal{N}_d(t)} = \underset{\boldsymbol{\Omega}(t_k)\succ 0, k\in\mathcal{N}_d(t)}{\arg\min} \frac{1}{\sqrt{|\mathcal{N}_d(t)|}}\sum_{k\in\mathcal{N}_d(t)}\left(\langle\boldsymbol{\Omega}(t_k),\tilde{\boldsymbol{H}}(t_k,h)\rangle_F - \mathrm{logdet}\boldsymbol{\Omega}(t_k)\right) \\ +\lambda\sum_{r\neq s} w_{rs}\left(\sum_{k\in\mathcal{N}_d(t)}\|\boldsymbol{\Omega}_{rs}(t_k)\|_F^2\right)^{1/2}.\end{aligned} \tag{3}$$

The group lasso penalty is also applied to each off-diagonal block of $\boldsymbol{\Omega}(t_k)$. Then, we introduce the following lemma to demonstrate that the reformulated problem is equivalent to the previous local group lasso estimator.

> **Lemma 1.** For any selection of h,d,λ, the solutions to problems (2) and (3), $\{\hat{\boldsymbol{\Theta}}(t_k;h,d,\lambda)\}_{k\in\mathcal{N}_d(t)}$ and $\{\hat{\boldsymbol{\Omega}}(t_k;h,d,\lambda)\}_{k\in\mathcal{N}_d(t)}$, respectively, must satisfy
>
> $$\hat{\boldsymbol{\Theta}}(t_k) = \begin{bmatrix} 1+[\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet}^{\top}\hat{\boldsymbol{\Omega}}(t_k)[\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet} & -[\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet}^{\top}\hat{\boldsymbol{\Omega}}(t_k) \\ -\hat{\boldsymbol{\Omega}}(t_k)[\tilde{\boldsymbol{\mu}}(t_k,h)]_{\bullet} & \hat{\boldsymbol{\Omega}}(t_k)\end{bmatrix},$$

for all $k\in\mathcal{N}_d(t)$.

From Lemma 1, we are able to discover that the reformulated problem solution shows the same sparsity pattern as that from the original problem, so we can condivert to solving the simplified problem to directly learn the graph topology. Also, the solution to the original problem can be recovered from $\{\hat{\boldsymbol{\Omega}}(t_k)\}_{k\in\mathcal{N}_d(t)}$. These results guarantee the equivalence between the two problems (2) and (3).

3 | COMPUTATION

3.1 | An alternating direction method of multipliers algorithm

The local group lasso estimator summarized by (3), as a solution to a convex problem, could be solved in the scheme of the two-block alternating direction method of multipliers (ADMM) (Boyd, 2010), which splits a complex problem into multiple subproblems through augmentations.

First, we notice that the problem can be rewritten into the following split form:

$$\min_{\boldsymbol{\Omega}(t_k),\boldsymbol{Z}(t_k),k\in\mathcal{N}_d(t)} \frac{1}{\sqrt{|\mathcal{N}_d(t)|}}\sum_{k\in\mathcal{N}_d(t)}\left(\langle\boldsymbol{\Omega}(t_k),\tilde{\boldsymbol{H}}(t_k,h)\rangle_F - \mathrm{logdet}\boldsymbol{\Omega}(t_k)\right)$$
$$+\lambda\sum_{r\neq s} w_{rs}\left(\sum_{k\in\mathcal{N}_d(t)}\|\boldsymbol{Z}_{rs}(t_k)\|_F^2\right)^{1/2}, \text{ subject to } \boldsymbol{\Omega}(t_k)=\boldsymbol{Z}(t_k)\succ 0 \text{ for all } k\in\mathcal{N}_d(t).$$

Then, the augmented Lagrangian in the scaled form is

$$\frac{1}{\sqrt{|\mathcal{N}_d(t)|}}\sum_{k\in\mathcal{N}_d(t)}\left(\langle\boldsymbol{\Omega}(t_k),\tilde{\boldsymbol{H}}(t_k,h)\rangle_F - \mathrm{logdet}\boldsymbol{\Omega}(t_k)\right)+\lambda\sum_{r\neq s} w_{rs}\left(\sum_{k\in\mathcal{N}_d(t)}\|\boldsymbol{Z}_{rs}(t_k)\|_F^2\right)^{1/2}$$
$$+\frac{\rho}{2}\sum_{k\in\mathcal{N}_d(t)}\|\boldsymbol{\Omega}(t_k)-\boldsymbol{Z}(t_k)+\boldsymbol{U}(t_k)\|_F^2,$$

where $\{\boldsymbol{U}(t_k)\}_{k\in\mathcal{N}_d(t)}$ are the scaled dual variables and $\rho>0$. The ADMM algorithm updates three blocks of variables $\{\boldsymbol{\Omega}(t_k)\}_{k\in\mathcal{N}_d(t)}, \{\boldsymbol{Z}(t_k)\}_{k\in\mathcal{N}_d(t)}$ and $\{\boldsymbol{U}(t_k)\}_{k\in\mathcal{N}_d(t)}$ alternately, so the following steps should be executed iteratively until convergence.

i. Update $\{\boldsymbol{\Omega}(t_k)\}_{k\in\mathcal{N}_d(t)}$. For each time point within the neighbourhood, do the eigen decomposition $\rho(\boldsymbol{Z}(t_k)-\boldsymbol{U}(t_k))-\left(1/\sqrt{|\mathcal{N}_d(t)|}\right)\tilde{\boldsymbol{H}}(t_k,h)=\mathbf{Q}(t_k)\boldsymbol{\Lambda}(t_k)\mathbf{Q}(t_k)^\top$, and then we update

$$\boldsymbol{\Omega}(t_k)\leftarrow\frac{1}{2\rho}\mathbf{Q}\left(\boldsymbol{\Lambda}+\sqrt{\boldsymbol{\Lambda}^2+\frac{4\rho}{\sqrt{|\mathcal{N}_d(t)|}}\boldsymbol{I}}\right)\mathbf{Q}^\top.$$

ii. Update $\{\boldsymbol{Z}(t_k)\}_{k\in\mathcal{N}_d(t)}$. We denote $\boldsymbol{\Omega}_{rs}(t_k)+\boldsymbol{U}_{rs}(t_k)$ by $\boldsymbol{A}_{rs}(t_k)$. For any $r\neq s$ and $k\in\mathcal{N}_d(t)$, the analytic solution becomes

$$\boldsymbol{Z}_{rs}(t_k)\leftarrow\left(1-\frac{\lambda w_{rs}}{\rho\sqrt{\sum_{g\in\mathcal{N}_d(t)}\|\boldsymbol{A}_{rs}(t_g)\|_F^2}}\right)_+\cdot\boldsymbol{A}_{rs}(t_k).$$

Also update all non-edge potentials, $\boldsymbol{Z}_{rr}(t_k)\leftarrow\boldsymbol{\Omega}_{rr}(t_k)+\boldsymbol{U}_{rr}(t_k)$.

iii. Update $\{\boldsymbol{U}(t_k)\}_{k\in\mathcal{N}_d(t)}$. Recalculate dual variables by $\boldsymbol{U}\leftarrow\boldsymbol{U}+\boldsymbol{\Omega}-\boldsymbol{Z}$.

To check the convergence of the ADMM algorithm, the standard stopping criterion on primal and dual residuals is adopted. Applying a moderate relative tolerance of 5×10^{-4} in our paper, satisfactory accuracy in recovery of sparsity pattern can be achieved within an acceptable number of iterations. In addition, the ADMM framework is also applicable to the model refitting procedure, where the $\boldsymbol{Z}$ update should be replaced by a projection to impose the equality constraint.

3.2 | Acceleration via exploiting block diagonal structure

In the previous algorithm, each ADMM iteration requires $|\mathcal{N}_d(t)|$ eigen decompositions on $q\times q$ matrices, so the per iteration complexity is $\mathcal{O}(q^3)$. The time complexity grows rapidly as the graph dimension increases, making the computation expensive for high-dimensional problems. To accelerate the computation, we borrow the idea from Danaher et al. (2014), Liu and Zhang (2020a) and Witten et al. (2011), which takes advantage of the block diagonal structure in the solution for the Gaussian graphical lasso problem. In this part, we propose a necessary and sufficient condition to determine whether the solution to problem (3) is block diagonal after some permutation of features. Given that condition, we are able to perform the previous ADMM algorithm within each block of features separately, which has the same solution as the one solved by applying the algorithm to all p features.

Lemma 2. Suppose that the solution to the alternative optimization problem (3) is block diagonal, that is, the p features can be reordered in some way such that the solution has the following block diagonal form:

$$\hat{\boldsymbol{\Omega}}(t_k) = \begin{bmatrix} \hat{\boldsymbol{\Omega}}_1(t_k) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\boldsymbol{\Omega}}_G(t_k) \end{bmatrix}.$$

Then, for any $g = 1,\ldots,G$, $\{\hat{\boldsymbol{\Omega}}_g(t_k)\}_{k\in\mathcal{N}_d(t)}$ can be obtained from solving the problem only on the corresponding g th set of variables.

Therefore, we are able to separate the local estimation problem into G subproblems according to the block diagonal structure. This leads to a massive reduction in computation complexity from $\mathcal{O}(q^3)$ to $\sum_{g=1}^{G}\mathcal{O}(q_g^3)$ per iteration, when we are able to divide the p features into a large number of blocks. To validate and identify the block diagonal structure of $\{\hat{\boldsymbol{\Omega}}(t_k)\}_{k\in\mathcal{N}_d(t)}$, we propose the following necessary and sufficient condition on the weighted empirical covariance matrices $\left\{\tilde{\boldsymbol{H}}(t_k,h)\right\}_{k\in\mathcal{N}_d(t)}$ for the presence of a block diagonal structure.

Theorem 1. *Let* C_1 *and* C_2 *be a non-overlapping partition of the* p *variables, such that* $C_1 \cap C_2 = \emptyset, C_1 \cup C_2 = \{1,\ldots,p\}$. *Then the following condition is necessary and sufficient for the variables in* C_1 *to be completely disconnected from those in* C_2 *in the estimated network:*

$$\frac{1}{|\mathcal{N}_d(t)|}\sum_{k\in\mathcal{N}_d(t)}\left\|\tilde{\boldsymbol{H}}_{rs}(t_k,h)\right\|_F^2 \le \lambda^2 w_{rs}^2 \text{ , for all } r\in C_1, s\in C_2.$$

Proof of these results can be found in the supporting information. It is noteworthy that this theorem can be easily extended to any number of unconnected blocks, so it allows us to check whether the solution to (3) has a block diagonal structure given a specific partition of the p variables. In this paper, for any fixed value of λ, we implement the following procedure to identify the block structure and the corresponding feature partition.

i. Create a $p \times p$ pseudo adjacency matrix $\boldsymbol{T}$ with $\boldsymbol{T}_{rr} = 1$ for $r = 1,\ldots,p$. For $r < s$, we check the condition in Theorem 1. If the condition is met, set $\boldsymbol{T}_{rs} = \boldsymbol{T}_{sr} = 0$. Otherwise, let $\boldsymbol{T}_{rs} = \boldsymbol{T}_{sr} = 1$.
ii. Use $\boldsymbol{T}$ to identify the connected and unconnected collections of variables and determine the partition to form the block diagonal structure. This step can be accomplished by the graph component searching algorithm proposed by Tarjan (1972), and the computational cost is no greater than $\mathcal{O}(p^2)$.

Please note that the pseudo adjacency matrix $\boldsymbol{T}$ is designated for detecting the block diagonal structure, and it does not represent the edge topology in most situations. When the tuning parameter λ is sufficiently large, the scalability of our algorithm can be substantially improved by divide and conquer. As for high-dimensional problems, only sparse networks with a small proportion of nodes being connected are useful and easy to interpret in real-world applications. So the proposed fast computation approach is helpful in most high-dimensional problems we encounter.

4 | SELECTION OF TUNING PARAMETERS

Our proposed local estimator is associated with three tuning parameters, namely, h, the kernel weight bandwidth; d, the neighbourhood width; and λ to control the overall sparsity. In this section, we introduce a V-fold time-stratified cross-validation (CV) scheme to select these tuning parameters, an approach that is adaptive to data with different levels of temporal smoothness. We denote $h_{\text{grid}}, d_{\text{grid}}, \lambda_{\text{grid}}$ the candidate tuning grids from which the parameters h, d, λ, respectively, should be chosen. Let $\mathcal{T}$ denote the collection of time points to be estimated. As recommended by Yang and Peng (2020), one plausible strategy is to choose d and λ separately for each $t\in\mathcal{T}$ due to the fact that the levels of sparsity and smoothness of the graph topology might be time varying, but a common h for all estimated time points.

To begin with, we create the validation datasets by including every Vth observed time point, so that the vth validation set includes an equally spaced sequence $\{\boldsymbol{X}(t_k) : k = v + cV, c\in\mathbb{Z}, k \le K\}$. The corresponding training set should consist of the remaining observed time points. For each $t\in\mathcal{T}$, we obtain its locally estimated model using only the training set data subject to a specific combination of tuning parameters. We denote the refitted solution at time t computed by the training set leaving out the vth fold by $\hat{\boldsymbol{\Theta}}^{\text{rf}}_{[-v]}(t;h,d,\lambda)$. The CV score measures how the fitted model predicts the adjusted empirical covariance matrix obtained from the validation set. In order to stabilize the CV result, we apply a kernel weighting on the validation set with the smallest bandwidth. Afterwards, the CV error for model prediction at time t on the vth validation set can be written as

$$\text{CV}_{[v]}(t;h,d,\lambda) = \left\langle \hat{\boldsymbol{\Theta}}^{\text{rf}}_{[-v]}(t;h,d,\lambda), \tilde{\boldsymbol{\Sigma}}_{[v]}(t,\min(h_{\text{grid}}))\right\rangle_F - \text{logdet}\left(\hat{\boldsymbol{\Theta}}^{\text{rf}}_{[-v]}(t;h,d,\lambda)\right),$$

which is the approximate log-likelihood computed by the trained solution and the validation data. This procedure is repeated V times. Then we take the arithmetic mean over the V folds and get $\mathrm{CV}(t;h,d,\lambda)$.

In principle, we would look for the minimum CV error to select the corresponding optimal combination of tuning parameters, which may be computationally expensive. To improve the efficiency, we suggest a line search over $h_{\mathrm{grid}}, d_{\mathrm{grid}}, \lambda_{\mathrm{grid}}$. First, we fix the smoothness parameters h and d at their median level, and we minimize the CV score with respect to λ separately for each estimated time point. Second, for each $t \in \mathcal{T}$, we fix the kernel bandwidth h at its median level and search the optimal neighbourhood width $\hat{d}(t)$ under the previously tuned $\hat{\lambda}(t)$. Ultimately, the optimal $\hat{h}$ is chosen by minimizing the sum of $\mathrm{CV}(t;h,\hat{d}(t),\hat{\lambda}(t))$ over those time points of interests. The effectiveness of our proposed method to select tuning parameters is verified in the following section through simulations.

5 | NUMERICAL EXPERIMENTS

5.1 | Setting

In this section, we generate synthetic data to illustrate the efficiency of the proposed local estimator as well as the validity of the scheme for tuning parameter selection. Networks gradually evolve over $K=101$ equally spaced time points $(0, 0.01, \dots, 1)$ in our experiment. Each network is a mixture of 10 categorical variables (3 categories), 10 truncated Poisson variables with truncation at 10 and 10 Gaussian variables, which we refer to as the "C–P–G mixture." As what we stated in the assumption, the underlying networks should be sparse and change smoothly. We let all node potential be time invariant, that is, $\boldsymbol{\theta}_r(t_k)=0$ for categorical and Gaussian variables; $\boldsymbol{\theta}_r(t_k)=-0.2$ for truncated Poisson variables; and $\boldsymbol{\theta}_{rr}(t_k)=-3$ for all Gaussian variables. The evolution of edge potentials are generated by the following models. For each vertex pair $r<s$, we separately generate random matrices $\boldsymbol{R}_1,\boldsymbol{R}_2,\boldsymbol{R}_3,\boldsymbol{R}_4 \in \mathbb{R}^{m_r \times m_s}$ with entries independently drawn from $\mathcal{N}(0, 0.5^2)$. Then we define

- Model 1: $\boldsymbol{G}_{rs}(t)=\sin(\pi t/2)\boldsymbol{R}_1+\cos(\pi t/2)\boldsymbol{R}_2+\sin(\pi t/4)\boldsymbol{R}_3+\cos(\pi t/4)\boldsymbol{R}_4$;
- Model 2: $\boldsymbol{G}_{rs}(t)=\sin(\pi t/4)\boldsymbol{R}_1+\cos(\pi t/4)\boldsymbol{R}_2+\sin(\pi t/8)\boldsymbol{R}_3+\cos(\pi t/8)\boldsymbol{R}_4$.

In order to add sparsity to edge potentials, we employ the following soft-thresholding:

$$\boldsymbol{G}'_{rs}(t)=\left(1-\frac{\sqrt{m_r m_s}}{\|\boldsymbol{G}_{rs}(t)\|_F}\right)_+ \cdot \boldsymbol{G}_{rs}(t).$$

We assign $\boldsymbol{\theta}_{rs}(t_k)=0.3\boldsymbol{G}'_{rs}(t_k)$ for all of the P–P, C–P and P–G edges; $\boldsymbol{\theta}_{rs}(t_k)=1.5\boldsymbol{G}'_{rs}(t_k)$ for the other types of edges. It results in an average of 68 and 61 edges in Model 1 and Model 2, respectively, around 15% of all vertex pairs. Based on the model formulation, Model 2 is supposed to have a smoother topology change in comparison with Model 1. As the model parameters gradually evolve, there exist many weak signals near the varying edges, which makes joint modelling crucial in topology estimation. Random samples are generated by Gibbs samplers under this setting with appropriate burn-in and thinning. The per time-point sample size is set at $n_k=20$. More details about the graph topology can be found in the supporting information.

We devote our simulation to the structural estimation at 6 time points $\mathcal{T}=\{0,0.2,\dots,1\}$. To obtain the weighted empirical covariance matrix, we adopt the unbounded Gaussian kernel

$$\mathcal{K}_h(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2h^2}\right).$$

Time-stratified fivefold CV is used to select the tuning parameters with the searching grids $h_{\mathrm{grid}}=\{0.1,0.2,\dots,0.5\}, d_{\mathrm{grid}}=\{0.01,0.03,0.05,0.1,0.15,0.2,0.25,0.3,0.4\}$ and $\lambda_{\mathrm{grid}}=\{10^{-2.5},10^{-2.4},\dots,10^{-1}\}$. Meanwhile, we also compare our proposed `LoALDIG` with the `tvmgm` function from the R package **mgm** (Haslbeck & Waldorp, 2020). The competing method is based on kernel smoothed neighbourhood regression, whose kernel bandwidth and sparsity are tuned by a combination of CV and extended Bayesian information criterion (EBIC). We treat the count variables as Poisson variables assuming that the true truncation points are unknown. Considering only two-way interactions in `tvmgm`, we employ the default "AND rule" to construct the topology estimates.

With regard to edge recovery, the problem of selecting edges can be treated as a binary classification problem. The metrics used for performance evaluation include the true positive rate (TPR) and the false discovery rate (FDR). Another popular measure for such problems is the F_1 score:

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \in [0,1],$$

which is an overall metric for model selection performance to balance TPR and FDR. The larger F_1 is, the better a method performs with respect to edge selection.

5.2 | Performance

First, we compare LoALDIG and tvmgm without consideration of tuning parameter selection to show the flexibility of LoALDIG. For each $t \in \mathcal{T}$, we utilize the two approaches to estimate its topology under all possible combinations of tuning parameters. Remark that the smoothness level of LoALDIG depends on hyperparameters h and d, but h is the only factor controlling the smoothness of tvmgm. To have fair comparison, we study how the F_1 score changes over the kernel bandwidth h in the two methods. Assuming the true topology is known, we search hyperparameters λ (and d) separately for each $t \in \mathcal{T}$ to optimize the F_1 score, so that we attain the highest F_1 score each method can achieve given each kernel bandwidth. Figure 1 depicts how the F_1 score varies over kernel bandwidth. All F_1 scores shown in the picture are averaged over 6 estimated time points and 20 independent replicates. Meanwhile, so as to better demonstrate the effect of neighbourhood width d on the performance of LoALDIG, we add three more lines standing for $d=0.01, d=0.2, d=0.4$ by using a common width on all estimated graphs.

Not surprisingly, our estimator reaches higher peaks than tvmgm thanks to exploiting the additional assumption on smooth change of sparsity structures. We also observe that tvmgm requires higher h to achieve the best edge selection performance compared with LoALDIG in Model 2. The reason is that a wider kernel can compensate for the lack of joint group lasso selection. Moreover, the lines of LoALDIG with optimal hyperparameter selection significantly outperform the ones using a common neighbourhood width for all time points to be estimated, which implies the necessity of separate tuning for d. In some cases such as $h=0.5, d=0.4$, overly wide neighbourhood sizes may diminish the temporal specificity, leading to the consequence that our estimate becomes not so good as the ones with narrower neighbourhoods. This phenomenon highlights the importance of proper tuning that is adaptive to different levels of smoothness from data. At last, we compare the edge selection performance under the two different true models. Our method enjoys better accuracy in inferring Model 2, since the smoother topological change from Model 2 is in favour of integrating information from nearby time points. Furthermore, the optimal F_1 score corresponds to higher kernel bandwidth in Model 2, which is also aligned with the reality that Model 2 has more temporal homogeneity.

Table 1 summarizes the simulation results tuned by time-stratified fivefold CV. Implemented on a laptop with 16-GB RAM and a 2.3-GHz dual-core Intel i5 processor, it takes less than 75 minutes on average to complete the CV. On average, the computation time of fitting one tuned model is around 10 seconds. For both Model 1 and Model 2, we observe that LoALDIG results in a higher F_1 score than tvmgm. Specifically, our proposed method recovers higher percentages of true edges, but yields more false positives at the same time. In spite of lower FDRs, the competing method is much too conservative, selecting fewer than 25% of true edges while missing most weak edge signals, since its model tuning is in favour of sparse models. In general, our method shows better balance between sensitivity and specificity. To obtain a closer overlook on the tuning parameter selection, we create Figure 2 to inspect the kernel bandwidth selected by fivefold CV. For both estimators, the CV selects a

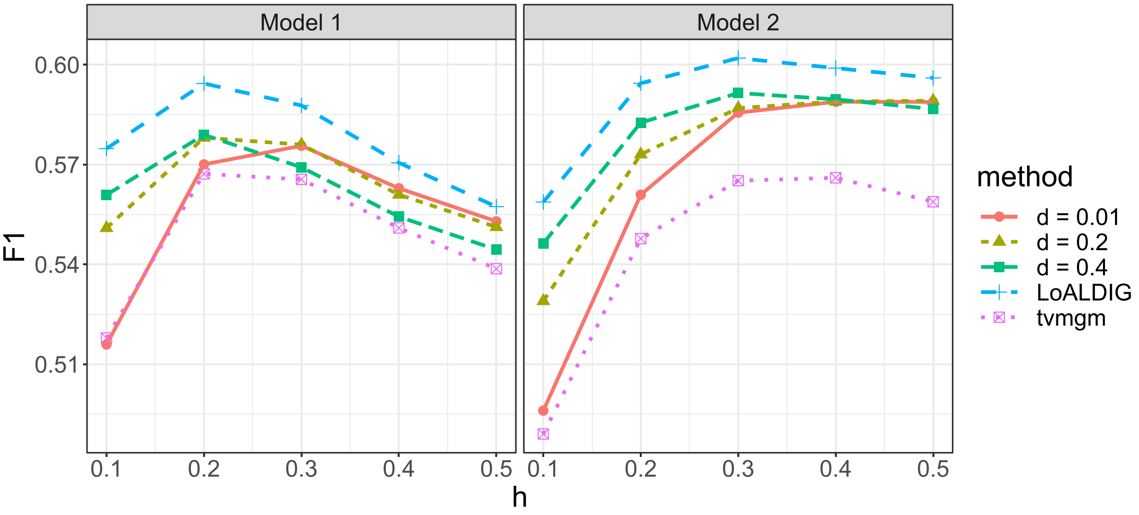

FIGURE 1 Comparison of edge recovery performance between LoALDIG and tvmgm: F_1 scores (averaged over 20 replicates) versus kernel bandwidth h

smoother kernel for Model 2, which is consistent with the true smoothness level. Specifically for `LoALDIG`, the time-stratified CV results in optimal kernel bandwidths as in Figure 1 for both models. More details about the simulation are provided in the supporting information.

6 | REAL DATA APPLICATION: TIME-VARYING HUMAN BRAIN GENE EXPRESSION–DNA METHYLATION NETWORKS

In this section, we conduct a case study on the PsychENCODE Human Brain Development dataset with free online access (https://development.psychencode.org). As the most complex organ in human body, brain is responsible for human's cognition, emotion and behaviour. Regulated by an enormously large number of factors, the process of human brain development is intricate and the full picture of regulatory mechanism is still to be uncovered. To have a deeper insight into the regulatory relationships among heterogeneous variables, mixed graphical models is a powerful tool, helping us better understand the functionalities of different types of factors and detect the dependency structures among them. Along with the PsychENCODE Human Brain Development dataset, Li et al. (2018) conducted a comprehensive analysis by integrating regulatory, epigenomic

TABLE 1 Simulation results: Model selection by fivefold CV

		TPR	FDR	F_1 score
Model 1	`LoALDIG`	0.519 (0.006)	0.385 (0.009)	0.562 (0.003)
	`tvmgm`	0.233 (0.012)	0.038 (0.010)	0.370 (0.015)
Model 2	`LoALDIG`	0.474 (0.009)	0.295 (0.011)	0.565 (0.006)
	`tvmgm`	0.223 (0.016)	0.021 (0.005)	0.357 (0.021)

Note: Means (and standard errors) are computed over 20 replicates.

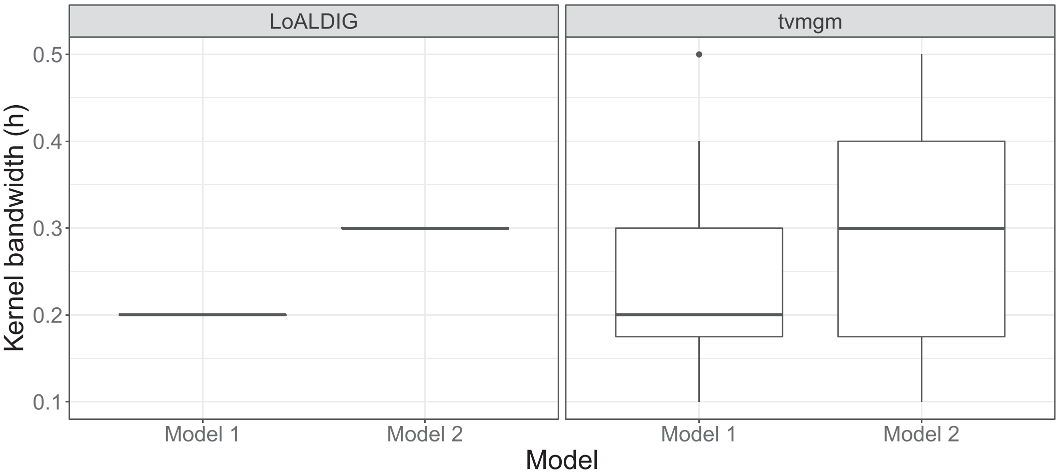

FIGURE 2 Boxplots of kernel bandwidth *h* selected by fivefold CV

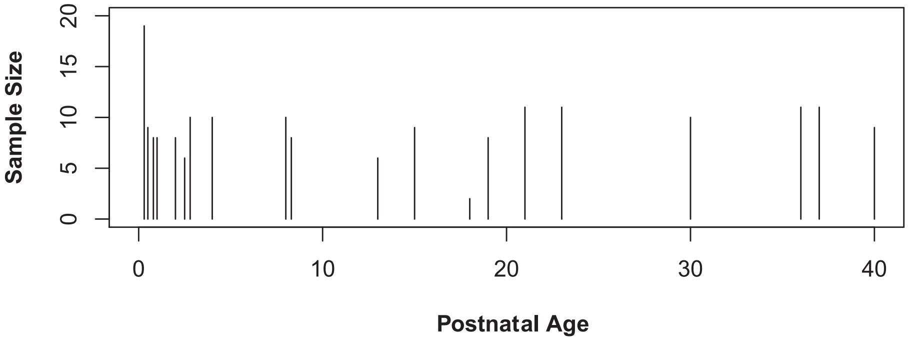

FIGURE 3 Number of observations along the developmental time span of human brain

and transcriptomic features of the human brain across time, regions and cell types. They discovered spatio-temporal trends and variations in brain development and related them to the major neuropsychiatric disorders.

Rather than the entire transcriptome, we focus our research on the the temporal correlation variation in a smaller co-expression module (the ME37 from Li et al., 2018). Specifically, we choose 26 important genes within this module, which are associated with multiple neurological disorders and traits such as schizophrenia, bipolar disorder and intelligence quotient (Li et al., 2018). In addition to gene expressions, we also consider DNA methylation in our study. As one of the most common epigenetic modifications, DNA cytosine methylation can regulate the transcription activity without changing the DNA sequence. Incorporating mRNA-seq and DNA methylation data along different stages of brain

TABLE 2 Summary of estimated gene expression–DNA methylation networks at four brain developmental stages

Stage	$\hat{d}(t)$	Total edges	RNA–RNA	RNA–methylation
1	0.05	90	57	33
2	0.4	79	46	33
3	0.1	124	88	36
4	0.03	17	15	2

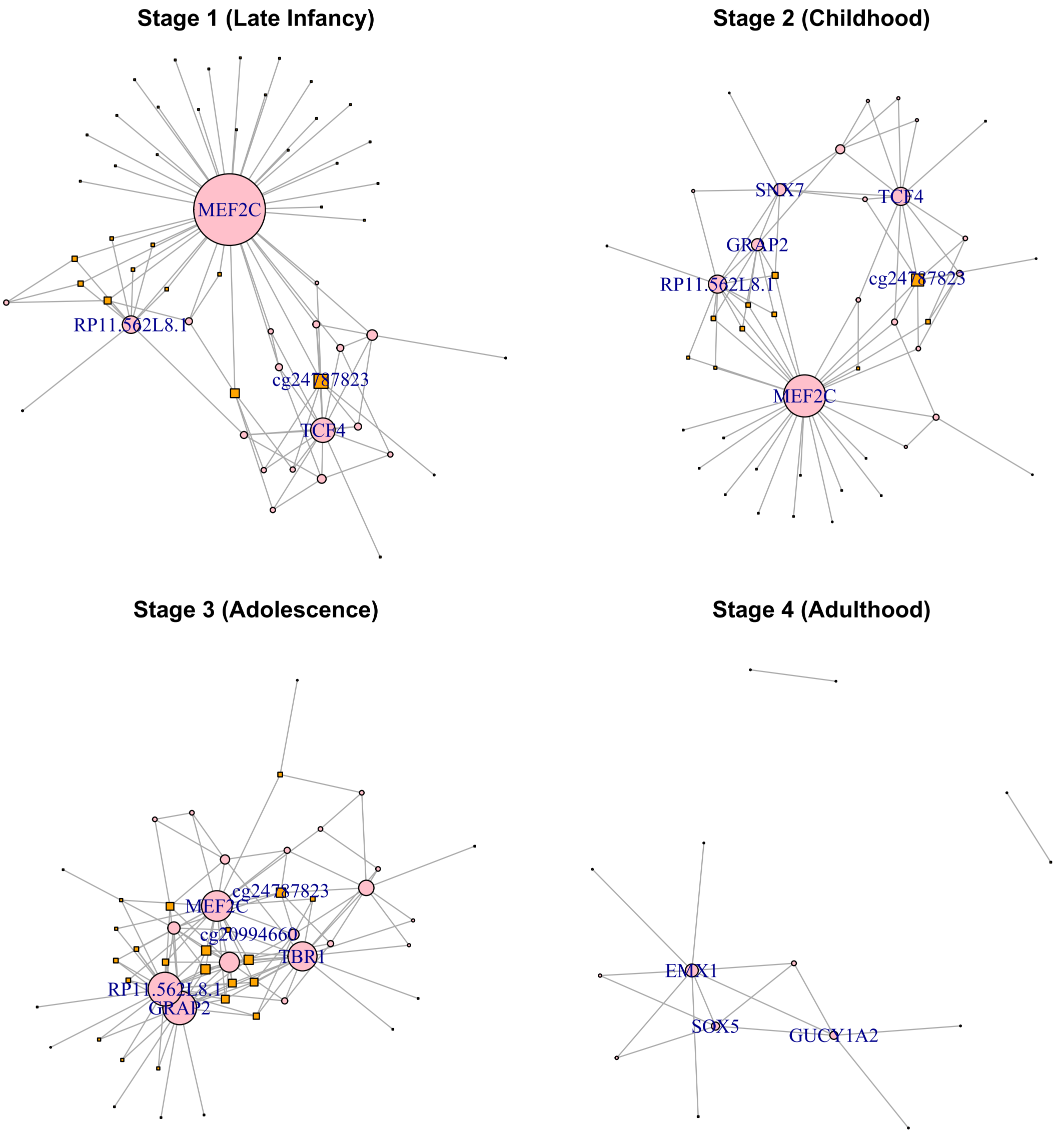

FIGURE 4 Time-varying gene expression–DNA methylation regulatory networks. Gene expressions are shown as pink circles; DNA methylations are presented as orange squares. The number of connections are node sizes

development to learn the time-varying mixed network structures, we are able to attain two major types of interactions of interest from our model, gene co-expression relationships as well as regulatory relationships between DNA methylation and gene expressions.

After data merging, we obtain a total of 21 subjects possessing both mRNA-seq and DNA methylation measurements, whose ages range from 4 postnatal months to 40 postnatal years. For each donor, tissues from different anatomical brain regions are collected. Since transcription profiles in postnatal neocortex (NCX) are relatively homogeneous (Kang et al., 2011), we select only tissue samples from 11 NCX areas into our analysis to ensure the homogeneity of samples. Although each individual donates multiple samples, we ignore the underlying dependency structure among the samples and treat them as independent observations. As a result, there are $K=20$ observed time points and the per time-point sample sizes fall between 2 and 19 (see Figure 3). The raw RNA-Seq count data are processed by the function `processSeq` from the R package **XMRF** (Allen & Liu, 2012), which adjusted for sequencing depth and overdispersion (Li et al., 2012). Then the processed mRNA-seq data are Poisson like, thereby modelled as count variables. As for DNA methylation, we incorporate CpG sites proximal to the transcription start sites (TSS) as well as the ones inside the 26 gene bodies and then filter out the sites whose methylation beta value standard deviations are lower than 0.05. Ultimately, a number of 94 CpG sites are selected. Their beta values are treated as Gaussian variables because of their shapes. The full list of genes and methylation sites can be found in the supporting information.

We are concerned about inferring the mixed networks at $\mathcal{T}=\{1,8,16,36\}$ years of age, representing late infancy, childhood, adolescence and adulthood, respectively. Each of them indicates a distinct stage of brain development (Kang et al., 2011). The optimal tuning parameters are selected by the aforementioned scheme of fivefold time-stratified CV, with search grids $h_{\text{grid}}=\{0.1,0.2,\ldots,0.5\}, d_{\text{grid}}=\{0.005,0.01,0.03,0.05,0.1,0.15,0.2,0.3,0.4\}$. As a result, we use $\hat{h}=0.3$ to infer the time-varying models, and the results are summarized in Table 2.

We recover the most edges in the adolescence stage and the fewest edges in the adulthood stage. No methylation–methylation interactions are found. From Figure 4, we can see the expression of MEF2C serves as a hub node in the first three stages, which agrees with the existing knowledge that MEF2C controls activity-dependent expression of neuronal genes, including those linked to the synapse function and autism spectrum disorder (Ebert & Greenberg, 2013; Parikshak et al., 2013). Other hub gene expression nodes, such as TCF4 and TBR1, also play critical roles in neurodevelopment (Huang & Hsueh, 2015). In the adolescence stage, the CpG site cg20994660 is the most connected DNA methylation, which is located at the TSS of gene TBR1. Based on the refitted model, the methylation level of cg20994660 is negatively correlated with the expression of TBR1, indicating the effect of transcriptional silencing.

7 | CONCLUDING REMARKS

In this paper, we have proposed a structural learning framework (`LoALDIG`) for estimating time-varying mixed graphical models. Through both kernel smoothing and a joint group lasso penalty, our local estimator fully exploits the temporal smoothness with much flexibility and therefore makes the graph estimation more accurate and efficient. Our method can complete the estimation of graphs of all time points simultaneously and avoid the common asymmetry problem in edge recovery. The practical performance has been validated through a numerical experiment and a real data application.

ACKNOWLEDGEMENT

Yuping Zhang gratefully acknowledges the National Science Foundation (NSF) grant Division of Mathematical Sciences (DMS) 2015481.

DATA AVAILABILITY STATEMENT

The following supporting information is available as part of the online article: code for the proposed `LoALDIG` method; and proofs, technical details and further information about simulation and real data application.

We conduct a case study on the PsychENCODE Human Brain Development dataset with free online access (http://development.psychencode.org).

ORCID

Qingyang Liu https://orcid.org/0000-0002-7811-2681
Yuping Zhang https://orcid.org/0000-0001-8986-0354
Zhengqing Ouyang https://orcid.org/0000-0003-2842-8503

REFERENCES

Allen, G. I., & Liu, Z. (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1–6.

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122. https://doi.org/10.1561/2200000016

Chautard, E., Thierry-Mieg, N., & Ricard-Blum, S. (2009). Interaction networks: From protein functions to drug discovery. A review. *Pathologie Biologie*, *57*(4), 324–333.

Chen, S., Witten, D. M., & Shojaie, A. (2014). Selection and estimation for mixed graphical models. *Biometrika*, *102*(1), 47–64.

Cheng, J., Li, T., Levina, E., & Zhu, J. (2017). High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, *26*(2), 367–378.

Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 373–397.

Ebert, D. H., & Greenberg, M. E. (2013). Activity-dependent neuronal signalling and autism spectrum disorder. *Nature*, *493*(7432), 327–337.

Gibberd, A. J., & Nelson, J. D. B. (2017). Regularized estimation of piecewise constant Gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, *26*(3), 623–634.

Gibberd, A. J., & Roy, S. (2017). Multiple changepoint estimation in high-dimensional Gaussian graphical models. arXiv preprint arXiv:1712.05786.

Hallac, D., Park, Y., Boyd, S., & Leskovec, J. (2017). Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 205–213.

Haslbeck, J., & Waldorp, L. (2020). mgm: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software, Articles*, *93*(8), 1–46. https://doi.org/10.18637/jss.v093.i08

Huang, T.-N., & Hsueh, Y.-P. (2015). Brain-specific transcriptional regulator T-brain-1 controls brain wiring and neuronal activity in autism spectrum disorders. *Frontiers in neuroscience*, *9*, 406.

Jordan, M. I., & Wainwright, M. J. (2004). Semidefinite relaxations for approximate inference on graphs with cycles. In *Advances in Neural Information Processing Systems*, pp. 369–376.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Ž., Mayer, S., Fertuzinhos, S., Umlauf, S. S., Lisgo, N., Vortmeyer, A., & Šestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483.

Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, *9*(10), 770.

Kolar, M., Song, L., Ahmed, A., & Xing, E. P. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, *4*(1), 94–123.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*(1), 31–57.

Lee, J. D., & Hastie, T. J. (2015). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, *24*(1), 230–253.

Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, *13*(3), 523–538.

Li, M., Santpere, G., Kawasawa, Y. I., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., Sunkin, S. M., Li, Z., Shin, Y., Zhu, Y., Sousa, A. M. M., Werling, D. M., Kitchen, R. R., Kang, H. J., Pletikos, M., Choi, J., Muchnik, S., Xu, X., Wang, D., Lorente-Galdos, B., & Sestan, N. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, *362*(6420), eaat7615.

Liu, Q., & Zhang, Y. (2020a). Fast variational inference for joint mixed sparse graphical models. *IEEE Journal on Selected Areas in Information Theory*, *1*(3), 908–913.

Liu, Q., & Zhang, Y. (2020b). Joint estimation of heterogeneous exponential Markov random fields through an approximate likelihood inference. *Journal of Statistical Planning and Inference*, *209*, 252–266.

Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., & Montana, G. (2014). Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, *103*, 427–443.

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., Horvath, S., & Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, *155*(5), 1008–1021.

Park, Y., Hallac, D., Boyd, S. P., & Leskovec, J. (2017). Learning the network structure of heterogeneous data via pairwise exponential Markov random fields.

Qiu, H., Han, F., Liu, H., & Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(2), 487–504.

Song, L., Kolar, M., & Xing, E. P. (2009). Keller: Estimating time-varying interactions between genes. *Bioinformatics*, *25*(12), i128–i136.

Tansey, W., Padilla, O. H. M., Suggala, A. S., & Ravikumar, P. (2015). Vector-space Markov random fields via exponential families. Icml.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing*, *1*(2), 146–160.

Wainwright, M. J., & Jordan, M. I. (2006). Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing*, *54*(6), 2099–2109.

Wainwright, M. J., & Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1* (1–2), 1–305.

Witten, D. M., Friedman, J. H., & Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, *20*(4), 892–900.

Yang, E., Allen, G., Liu, Z., & Ravikumar, P. K. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pp. 1358–1366.

Yang, E., Baker, Y., Ravikumar, P., Allen, G., & Liu, Z. (2014). Mixed graphical models via exponential families, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Reykjavik, Iceland: PMLR, pp. 1042–1050.

Yang, E., Ravikumar, P. K., Allen, G. I., & Liu, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems*, pp. 1718–1726.

Yang, J., & Peng, J. (2020). Estimating time-varying graphical models. *Journal of Computational and Graphical Statistics*, *29*(1), 191–202.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1), 17.

Zhang, Y., Ouyang, Z., & Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics*, *11*(1), 161–184.

Zhou, S., Lafferty, J., & Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, *80*(2-3), 295–319.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Liu, Q., Zhang, Y., & Ouyang, Z. (2021). Structural inference of time-varying mixed graphical models. *Stat*, *10*(1), e414. https://doi.org/10.1002/sta4.414