Semi-parametric Methods for Estimating Time-varying Graph Structure *

Mladen Kolar
mladenk@cs.cmu.edu
Machine Learning Department
School of Computer Science

January 24, 2010

Abstract

Stochastic networks are a plausible representation of the relational information among entities in dynamic systems such as living cells or social communities. While there is a rich literature in estimating a static or temporally invariant network from observation data, little has been done towards estimating time-varying networks from time series of entity attributes. In this paper, we present two new machine learning methods for estimating time-varying networks, which both build on a temporally smoothed l_1-regularized logistic regression formalism that can be cast as standard convex-optimization problem and solved efficiently using generic solvers scalable to large networks. We report promising results on recovering simulated time-varying networks. For real datasets, we reverse engineer the latent sequence of temporally rewiring political networks between Senators from the US Senate voting records and the latent evolving regulatory networks underlying 588 genes across the life cycle of *Drosophila melanogaster* from microarray time course. We provide some theoretical guarantees for the proposed methods.

*This work was done under supervision of my advisor Eric Xing. A part of this work is going to appear in Annals of Applied Statistics (Kolar, Song, Ahmed, Xing. Estimating Time-varying Networks). I am very grateful for multiple discussions I had with Larry Wasserman and John Lafferty.

Contents

1 Introduction

Consider the following real world problems:

- *Analysis of gene regulatory networks.* Suppose that we have a set of n microarray measurements of gene expression levels, obtained at different stages during the development of an organism or at different times during the cell cycle. Given this data, biologists would like to get insight into dynamic relationships between different genes and how these relations change at different stages of development. The problem is that at each time point there is only one or at most a few measurements of the gene expressions; and a naive approach to estimating the gene regulatory network, which uses only the data at the time point in question to infer the network, would fail. To obtain a good estimate of the regulatory network at any time point, we need to leverage the data collected at other time points and extract some information from them.

- *Analysis of stock market.* In a finance setting, we have values of different stocks at each time point. Suppose, for simplicity, that we only measure whether the value of a particular stock is going up or down. We would like to find the underlying transient relational patterns between different stocks from these measurements and get insight into how do these patterns change over time. Again, we only have one measurement at each time point and we need to leverage information from the data obtained at nearby time points.

- *Understanding social networks.* There are 100 Senators in the U.S. Senate and each can cast a vote on different bills. Suppose that we are given n voting records over some period of time. How can one infer the latent political liaisons and coalitions among different senators and the way these relationships change with respect to time and with respect to different issues raised in bills just from the voting records?

What is common to the above described problems is that they all concern with estimating a sequence of time-specific latent relational structures between a fixed set of entities (i.e., variables), from a time series of observation data of entities states; and the relational structures between the entities are time evolving, rather than being invariant throughout the data collection period as commonly assumed in nearly all previous work on structure estimation such as [3, 21, 24, 23]. Typically, the available data for the problem are very scarce, with only one or at most a few measurements per time point corresponding to any particular latent structure; and the data are very high-dimensional, with the total number of observations small compared to the total number of potential relations, which make the problem of structure estimation even more challenging than the static case studied recently by [23].

A popular model for the relational structure over a fixed set of entities that is widely studied is the Markov random field (MRF) [31, 12]. Let $G = (V, E)$

represent a graph, of which V denotes the set of vertices, and E denotes the set of edges over vertices. Depending on the specific application of interest, a node $u \in V$ can represent a gene, a stock, or a social actor, and an edge $(u, v) \in E$ can represent a relationship (e.g., correlation, influence, friendship) between actors u and v. Let $\mathbf{X} = (X_1, \ldots, X_p)'$, where $p = |V|$, be a random vector of nodal states following a probability distribution indexed by $\boldsymbol{\theta} \in \Theta$. Under a MRF, the nodal states X_u's are assumed to be discrete, i.e., $X_u \in \mathcal{X} \equiv \{s_1, \ldots, s_k\}$, and the edge set $E \subseteq V \times V$ encodes certain conditional independence assumptions among components of the random vector $\mathbf{X}$, for example, the random variable X_u is conditionally independent of the random variable X_v given the rest of the variables if $(u, v) \notin E$. Under the special case of binary nodal states, e.g., $X_u \in \mathcal{X} \equiv \{-1, 1\}$, and assuming pairwise potential weighted by θ_{uv} for all $(u, v) \in E$ and $\theta_{uv} = 0$ for all $(u, v) \notin E$, the joint probability of $\mathbf{X} = \mathbf{x}$ can be expressed by a simple exponential family model: $\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{Z} \exp\{\sum_{u<v} \theta_{uv} x_u x_v\}$, also known as the Ising model, where Z denotes the partition that is usually intractable to compute. A number of recent papers have studied in depth how to estimate this model from data that are assumed to be *i.i.d.* samples from the model, and the asymptotic guarantee of the estimator [23, 3]. In particular, an important focus has been on the problem of *structure estimation* of the graph topology represented by E. It has been shown that under certain variable conditions, it is possible to obtain an estimator of the edge set E that achieve a property known as *sparsistency* [23], which refers to the case where a consistent estimator of E can be attained when the true degree (i.e., number of neighbors) of each node is much smaller than the size of the graph p.

In this paper, we are interested in learning the graph structures of MRFs from observational data, but under a more demanding scenario where the data $\{\mathbf{x}^t\}$ are not *i.i.d.* samples from a time-invariant MRF, but from a series of time-evolving MRFs $\{\mathbb{P}_{\boldsymbol{\theta}^t}(\cdot)\}_{t \in \mathcal{T}_n}$, where $\mathcal{T}_n = \{1/n, 2/n, \ldots, 1\}$ is the time index set; and our goal is to estimate the sequence of graphs $\{G^t\}_{t \in \mathcal{T}_n}$ underlying each observation $\mathbf{x}^t \sim \mathbb{P}_{\boldsymbol{\theta}^t}$ in the time series, rather than a single static graph G underlying $\mathbb{P}_{\boldsymbol{\theta}}$ over all time points. Under the traditional assumption of data sampled *i.i.d.* from an invariant $\mathbb{P}_{\boldsymbol{\theta}}$, structural estimation of a MRF can be cast as a *neighborhood selection* problem for each node in the graph based on a ℓ_1-norm regularized regression procedure, of which the theoretical guarantees have been recently thoroughly studied [23], as we review shortly. We instead focus on estimating the graph structures from a set of n independent, high-dimensional observations which are NOT identically distributed, which is arguably a more realistic characteristic of the data. Because of this more general problem we are near the extremum of the high-p/low-n scenario for high-dimensional inference in the traditional sense, (i.e., n is approaching 1, corresponding to as few as 1 instance of $\mathbf{x}$ per time-specific MRF), it is intriguing to ask, can we reliably estimate the changing graph structure and, if so, under what conditions? It might seem that the problem is ill-defined, since for any time point we have at most one observation; however, as we will show shortly, under a set of suitable assumptions the problem is indeed well defined and the series of underlying graph structures can be estimated. For example, we may assume that the probability

distributions are changing *smoothly* over time, or there exist a partition of the interval $[0, 1]$ into segments where the graph structure within each segment is invariant.

It is noteworthy that the problem of the graph structure estimation is quite different from the problem of (value-) consistent estimation of the unknown parameter $\boldsymbol{\theta}$ that indexes the distribution. In general, the graph structure estimation requires a more stringent assumptions on the underlying distribution and the parameter values. For example, observe that a consistent estimator of $\boldsymbol{\theta}$ in the Euclidean distance does not guarantee a consistent estimation of the graph structure, encoded by the non-zero patter of the estimator. In the motivating problems that we started with, the main goal is to understand the interactions between different actors. These interactions are more easily interpreted by a domain expert than the numerical values of the parameter vector $\boldsymbol{\theta}$ and have potential to reveal more information about the underlying process of interest. This is especially true in situations where there is little or no domain knowledge and one is interested in obtaining casual, preliminary information. Furthermore, the problem of dynamic structure estimation is of high importance in domains that lack prior knowledge or measurement techniques about the interactions between different actors; and such estimates can provide desirable information about the details of relational changes in a complex system.

1.1 Related work

A large body of literature has focused on estimation of the time-invariant graph structure from the *i.i.d.* sample. Assume that $\mathcal{D}_n = \{\mathbf{x}^i = (x_1^i, \ldots, x_p^i)\}_{i=1}^n$ are n *i.i.d.* samples from $\mathbb{P}_{\boldsymbol{\theta}}$. Furthermore, under the assumption that $\mathbb{P}_{\boldsymbol{\theta}}$ is a multivariate normal distribution with mean vector μ and covariance matrix $\boldsymbol{\Sigma}$, estimation of the graph structure is equivalent to the estimation of zeros in the concentration matrix $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$ [19]. [5] proposed a method that tests if partial correlations are different from zero, which can be applied when the number of dimensions p is small in comparison to the sample size n. In the recent years, research has been directed towards methods that can handle datasets with relatively few high-dimensional samples, which are common if a number of domains, e.g., microarray measurement experiments, fMRI datasets and astronomical measurements. These "large p, small n" datasets pose a difficult estimation problem, but under the assumption that the underlying graph structure is sparse, several methods can be employed successfully for structure recovery. [21] proposed a procedure based on *neighborhood selection* of each node via the ℓ_1 penalized regression. This procedure uses a pseudo-likelihood, which decomposes across different nodes, to estimate graph edges and, although the estimated parameters are not consistent, the procedure recovers the graph structure consistently under a set of suitable conditions. A related approach is proposed in [22] who consider a different neighborhood selection procedure for the structure estimation in which they estimate all neighborhoods jointly and as a result obtain a global estimate of the graph structure that empirically improves the performance on a number of networks. These neighborhood selec-

tion procedures are suitable for large-scale problems due to availability of fast solvers to ℓ_1 penalized problems [7, 10].

Another popular approach to the graph structure estimation is the ℓ_1 penalized likelihood maximization, which simultaneously estimates the graph structure and the elements of the covariance matrix, however, at a price of computational efficiency. The penalized likelihood approach involves solving a semidefinite program (SDP) and a number of authors have worked on efficient solvers that exploit the special structure of the problem [2, 33, 11, 6, 26]. Of these methods, it seems that the graphical lasso [11] is the most computationally efficient. Some authors have proposed to use a non-concave penalty instead of the ℓ_1 penalty, which tries to remedy the bias that the ℓ_1 penalty introduces [16, 8, 35].

When the random variable $\mathbf{X}$ is discrete, the problem of structure estimation becomes even more difficult since the likelihood cannot be optimized efficiently due to the intractability of evaluation of the log-partition function. [23] use a pseudo-likelihood approach, based on the local conditional likelihood at each node, to estimate the neighborhood of each node, and show that this procedure estimates the graph structure consistently.

All of the aforementioned work analyzes estimation of a time-invariant graph structure from an *i.i.d.* sample. On the other hand, with few exceptions [15, 27, 14, 34], much less has been done on modeling dynamical processes that guide topological rewiring and semantic evolution of networks over time. In particular, very little has been done towards estimating the time-varying graph topologies from observed nodal states, which represent attributes of entities forming a network. [15] introduced a new class of models to capture dynamics of networks evolving over discrete time steps, called *temporal Exponential Random Graph Models* (tERGMs). This class of models uses a number of statistics defined on time-adjacent graphs, e.g., "edge-stability," "reciprocity," "density," "transitivity," etc., to construct a log-linear graph transition model $P(G^t|G^{t-1})$ that captures dynamics of topological changes. [14] incorporate a hidden Markov process into the tERGMs, which imposes stochastic constraints on topological changes in graphs, and, in principle, show how to infer a time-specific graph structure from the posterior distribution of G^t, given the time series of node attributes. Unfortunately, even though this class of model is very expressive, the sampling algorithm for posterior inference scales only to small graphs with tens of nodes.

The work of [34] is the most relevant to our work and we briefly describe it below. The authors develop a nonparametric method for estimation of time-varying Gaussian graphical model, under the assumption that the observations $\mathbf{x}^t \sim \mathcal{N}(0, \mathbf{\Sigma}^t)$ are independent, but not identically distributed, realizations of a multivariate distribution whose covariance matrix changes smoothly over time. The time-varying Gaussian graphical model is a continuous counterpart of the discrete Ising model considered in this paper. In [34], the authors address the issue of consistent, in the Frobenius norm, estimation of the covariance and concentration matrix, however, the problem of consistent estimation of the non-zero pattern in the concentration matrix, which corresponds to the graph

structure estimation, is not addressed there. Note that the consistency of the graph structure recovery does not immediately follow from the consistency of the concentration matrix.

The paper is organized as follows. In Section 2 we describe the proposed models for estimation of the time varying graphical structures and the algorithms for obtaining the estimators. In Section 3, the performance of the methods is demonstrated through simulation studies. In Section 4, the methods are applied to some real world data sets. In Section 5, we give theoretical properties of the algorithms. Discussion is given in Section 6.

2 Methods

Let $\mathcal{D}_n = \{\mathbf{x}^t \sim \mathbb{P}_{\boldsymbol{\theta}^t} | t \in \mathcal{T}_n\}$ be an independent sample of n observation from a time series, obtained at discrete time steps indexed by $\mathcal{T}_n = \{1/n, 2/n, \ldots, 1\}$ (for simplicity we assume that the observations are equidistant in time). Each sample point comes from a different discrete time step and is distributed according to a distribution $\mathbb{P}_{\boldsymbol{\theta}^t}$ indexed by $\boldsymbol{\theta}^t \in \Theta$. In particular, we will assume that $\mathbf{X}^t$ is a p-dimensional random variable taking values from $\{-1, 1\}^p$ with a distribution of the following form:

$$\mathbb{P}_{\boldsymbol{\theta}^t}(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\theta}^t)} \exp\left(\sum_{(u,v)\in E^t} \theta^t_{uv} x_u x_v\right), \tag{1}$$

where $Z(\boldsymbol{\theta}^t)$ is the partition function, $\boldsymbol{\theta}^t \in \mathbb{R}^{\binom{p}{2}}$ is the parameter vector and $G^t = (V, E^t)$ is an undirected graph representing conditional independence assumptions among subsets of the p-dimensional random vector $\mathbf{X}^t$. Recall that $V = \{1, \ldots, p\}$ is the node set and each node corresponds with one component of the vector $\mathbf{X}^t$. In the paper we are addressing the problem of graph structure estimation from the observational data which we now formally define: *given any time point $\tau \in [0, 1]$ estimate the graph structure associated with $\mathbb{P}_{\boldsymbol{\theta}^t}$, given the observations $\mathcal{D}_n$.* To obtain insight into the dynamics of changes in the graph structure one only needs to estimate graph structure for multiple time-point, e.g., for every $\tau \in \mathcal{T}_n$.

The graph structure G^τ is encoded by the locations of the non-zero elements of the parameter vector $\boldsymbol{\theta}^\tau$, which we refer to as the non-zero pattern of the parameter $\boldsymbol{\theta}^\tau$. Components of the vector $\boldsymbol{\theta}^\tau$ are indexed by distinct pairs of nodes and a component of the vector θ^τ_{uv} is non-zero if and only if the corresponding edge $(u, v) \in E^\tau$. Throughout the rest of the paper we will focus on estimation of the non-zero pattern of the vector $\boldsymbol{\theta}^\tau$ as a way to estimate the graph structure. Let $\boldsymbol{\theta}^\tau_u$ be the $(p-1)$-dimensional subvector of parameters

$$\boldsymbol{\theta}^\tau_u := \{\theta^\tau_{uv} \mid v \in V \backslash u\}$$

associated with each node $u \in V$, and let $S^\tau(u)$ be the set of edges adjacent to a node u at a time point τ:

$$S^\tau(u) := \{(u, v) \in V \times V \mid \theta^\tau_{uv} \neq 0\}.$$

Observe that the graph structure G^τ can be recovered from the local information on neighboring edges $S^\tau(u)$, for each node $u \in V$, which can be obtained from the non-zero pattern of the subvector $\boldsymbol{\theta}_u^\tau$ alone. The main focus of this section is on obtaining node-wise estimators $\hat{\boldsymbol{\theta}}_u^\tau$ of the non-zero pattern of the subvector $\boldsymbol{\theta}_u^\tau$, which are then used to create estimates

$$\hat{S}^\tau(u) := \{(u,v) \in V \times V \mid \hat{\theta}_{uv}^\tau \neq 0\}, \quad u \in V. \tag{2}$$

Note that the estimated non-zero pattern might be asymmetric, e.g., $\hat{\theta}_{uv}^\tau = 0$, but $\hat{\theta}_{vu}^\tau \neq 0$. We consider using the min and max operations to combine the estimators $\hat{\theta}_{uv}^\tau$ and $\hat{\theta}_{vu}^\tau$. Let $\tilde{\boldsymbol{\theta}}^\tau$ denote the combined estimator. The estimator combined using the min operation has the following form:

$$\tilde{\theta}_{uv} = \begin{cases} \hat{\theta}_{uv} & \text{if } |\hat{\theta}_{uv}| < |\hat{\theta}_{vu}| \\ \hat{\theta}_{vu} & \text{if } |\hat{\theta}_{uv}| \geq |\hat{\theta}_{vu}| \end{cases} \quad \text{“min\_symmetrization”}, \tag{3}$$

which means that the edge (u,v) is included in the graph estimate only if it appears in both estimates $\hat{S}^\tau(u)$ and $\hat{S}^\tau(v)$. Using the max operation, the combined estimator can be expressed as:

$$\tilde{\theta}_{uv} = \begin{cases} \hat{\theta}_{uv} & \text{if } |\hat{\theta}_{uv}| > |\hat{\theta}_{vu}| \\ \hat{\theta}_{vu} & \text{if } |\hat{\theta}_{uv}| \leq |\hat{\theta}_{vu}| \end{cases} \quad \text{“max\_symmetrization”}, \tag{4}$$

and as a result the edge (u,v) is included in the graph estimate if it appears in at least one of the estimate $\hat{S}^\tau(u)$ or $\hat{S}^\tau(v)$.

An estimator $\hat{\boldsymbol{\theta}}_u^\tau$ is obtained through the use of pseudo-likelihood based on the conditional distribution of X_u^τ given the other of variables $\mathbf{X}_{\setminus u}^\tau = \{X_v^\tau \mid v \in V \setminus u\}$. Although the use of pseudo-likelihood fails in certain scenarios, e.g., estimation of Exponential Random Graphs (see [30] for a recent study), the graph structure of an Ising model can be recovered from an *i.i.d.* sample using the pseudo-likelihood, as shown in [23]. Under the model (1), the conditional distribution of X_u^τ given the other variables $\mathbf{X}_{\setminus u}^\tau$ takes the form:

$$\mathbb{P}_{\boldsymbol{\theta}_u^\tau}(x_u^\tau | \mathbf{X}_{\setminus u}^\tau = \mathbf{x}_{\setminus u}^\tau) = \frac{\exp(x_u^\tau \langle \boldsymbol{\theta}_u^\tau, \mathbf{x}_{\setminus u}^\tau \rangle)}{\exp(x_u^\tau \langle \boldsymbol{\theta}_u^\tau, \mathbf{x}_{\setminus u}^\tau \rangle) + \exp(-x_u^\tau \langle \boldsymbol{\theta}_u^\tau, \mathbf{x}_{\setminus u}^\tau \rangle)}, \tag{5}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b}$ denotes the dot product. For simplicity, we will write $\mathbb{P}_{\boldsymbol{\theta}_u^\tau}(x_u^\tau | \mathbf{X}_{\setminus u}^\tau = \mathbf{x}_{\setminus u}^\tau)$ as $\mathbb{P}_{\boldsymbol{\theta}_u^\tau}(x_u^\tau | \mathbf{x}_{\setminus u}^\tau)$. Observe that the model given in Eq. (5) can be viewed as expressing X_u^τ as the response variable in the generalized varying-coefficient models with $\mathbf{X}_{\setminus u}^\tau$ playing the role of covariates. Under the model given in Eq. (5), the conditional log-likelihood, for the node u at the time point $t \in \mathcal{T}_n$, can be written in the following form:

$$\begin{aligned} \gamma(\boldsymbol{\theta}_u; \mathbf{x}^t) &= \log \mathbb{P}_{\boldsymbol{\theta}_u}(x_u^t | \mathbf{x}_{\setminus u}^t) \\ &= x_u^t \langle \boldsymbol{\theta}_u, \mathbf{x}_{\setminus u}^t \rangle - \log \left(\exp(\langle \boldsymbol{\theta}_u, \mathbf{x}_{\setminus u}^t \rangle) + \exp(-\langle \boldsymbol{\theta}_u, \mathbf{x}_{\setminus u}^t \rangle) \right). \end{aligned} \tag{6}$$

The non-zero pattern of $\boldsymbol{\theta}_u^\tau$ can be estimated by maximizing the conditional log-likelihood given in Eq. (6). What is left to show is how to combine the information across different time points, which will depend on the assumptions that are made on the unknown vector $\boldsymbol{\theta}^t$.

The primary focus is to develop methods applicable to datasets with the total number of observations n small compared to the dimensionality $p = p_n$. Without assuming anything about $\boldsymbol{\theta}^t$, the estimation problem is ill-posed, since there can be more parameters than samples. A common way to deal with the estimation problem is to assume that the graphs $\{G^t\}_{t\in\mathcal{T}_n}$ are sparse, i.e., the parameter vectors $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$ have only few non-zero elements. In particular, we assume that each node u has a small number of neighbors, i.e., there exist a number $s \ll p$ such that it upper bounds the number of edges $|S^\tau(u)|$ for all $u \in V$ and $\tau \in \mathcal{T}_n$. In many real data sets the sparsity assumption holds quite well. For example, in a genetic network, rarely a regulator gene would control more than a handful of regulatees under a specific condition [4]. Furthermore, we will assume that the parameter vector $\boldsymbol{\theta}^t$ behave "nicely" as a function of time. Intuitively, without any assumptions about the parameter $\boldsymbol{\theta}^t$ it is impossible to aggregate information from observations even close in time, because the underlying probability distributions for observations from different time points might be completely different. In the paper we will consider two ways of constraining the parameter vector $\boldsymbol{\theta}^t$ as a function of time:

- *Smooth changes in parameters.* We first consider that the distribution generating the observation changes smoothly over the time, i.e., the parameter vector $\boldsymbol{\theta}^t$ is a smooth function of time. Formally, we assume that there exists a constant $M > 0$ such that it upper bounds the following quantities:

$$\max_{u,v\in V\times V} \sup_{t\in[0,1]} |\frac{\partial}{\partial t}\theta_{uv}^t| < M, \qquad \max_{u,v\in V\times V} \sup_{t\in[0,1]} |\frac{\partial^2}{\partial t^2}\theta_{uv}^t| < M.$$

  Under this assumption, as we get more and more data (i.e. we collect data in higher and higher temporal resolution within interval $[0, 1]$), parameters, and graph structures, corresponding to any two adjacent time points will differ less and less.

- *Piecewise constant with abrupt structural changes in parameters.* Next, we consider that there are a number of change points at which the distribution generating samples changes abruptly. Formally, we assume that for each node u, there is a partition $\mathcal{B}_u = \{0 = B_{u,0} < B_{u,1} < \ldots < B_{u,k_u} = 1\}$ of the interval $[0, 1]$, such that each element of $\boldsymbol{\theta}_u^t$ is constant on each segment of the partition. At change points some of the elements of the vector $\boldsymbol{\theta}_u^t$ may become zero, while some others may become non-zero, which corresponds to a change in the graph structure. If the number of change points is small, i.e., the graph structure changes infrequently, then there will be enough samples at a segment of the partition to estimate the non-zero pattern of the vector $\boldsymbol{\theta}^\tau$.

In the following two subsections we propose two estimation methods, each suitable for one of the assumptions discussed above.

2.1 Smooth changes in parameters

Under the assumption that the elements of $\boldsymbol{\theta}^t$ are smooth functions of time, as described in the previous section, we use a kernel smoothing approach to estimate the non-zero pattern of $\boldsymbol{\theta}_u^\tau$ at the time point of interest $\tau \in [0,1]$, for each node $u \in V$. These node-wise estimators are then combined using either Eq. (3) or Eq. (4) to obtain the estimator of the non-zero pattern of $\boldsymbol{\theta}^\tau$. The estimator $\hat{\boldsymbol{\theta}}_u^\tau$ is defined as a minimizer of the following objective:

$$\hat{\boldsymbol{\theta}}_u^\tau := \min_{\boldsymbol{\theta}_u \in \mathbb{R}^{p-1}} \left\{ l\left(\boldsymbol{\theta}_u; \mathcal{D}_n\right) + \lambda_1 ||\boldsymbol{\theta}_u||_1 \right\} \tag{7}$$

where

$$l(\boldsymbol{\theta}_u; \mathcal{D}_n) = -\sum_{t \in \mathcal{T}_n} w_t^\tau \gamma(\boldsymbol{\theta}_u; \mathbf{x}^t) \tag{8}$$

is a weighted log-likelihood, with weights defined as $w_t^\tau = \frac{K_h(t-\tau)}{\sum_{t' \in \mathcal{T}_n} K_h(t'-\tau)}$ and $K_h(\cdot) = K(\cdot/h)$ is a symmetric, nonnegative kernel function. We will refer to this approach of obtaining an estimator as `smooth`. The ℓ_1 norm of the parameter is used to regularize the solution and as a result the estimated parameter has a lot of zeros. The number of the non-zero elements of $\hat{\boldsymbol{\theta}}_u^\tau$ is controlled by the user-specified regularization parameter $\lambda_1 \geq 0$. The bandwidth parameter h is also a user defined parameter that effectively controls the number of observations around τ used to obtain $\hat{\boldsymbol{\theta}}_u^\tau$. In Section 2.4 we discuss how to choose the parameters λ_1 and h.

The optimization problem (7) is the well known objective of the ℓ_1 penalized logistic regression and there are many ways of solving it, e.g., the interior point method of [17], the projected subgradient descent method of [6] or the fast coordinate-wise descent method of [9]. From our limited experience, the specialized first order methods work faster than the interior point methods and we briefly describe the iterative coordinate-wise descent method:

1. Set initial values: $\hat{\boldsymbol{\theta}}_u^{\tau,0} \leftarrow \mathbf{0}$

2. For each $v \in V \backslash u$, set the current estimate $\hat{\theta}_{uv}^{\tau,iter+1}$ as a solution to the following optimization procedure:

$$\min_{\theta \in \mathbb{R}} \left\{ \begin{array}{c} \sum_{t \in \mathcal{T}_n} \gamma\left(\hat{\theta}_{u,1}^{\tau,iter+1}, \ldots, \hat{\theta}_{u,v-1}^{\tau,iter+1}, \theta, \hat{\theta}_{u,v+1}^{\tau,iter}, \ldots, \hat{\theta}_{u,p-1}^{\tau,iter}; \mathbf{x}^t\right) \\ +\lambda_1 |\theta| \end{array} \right\}. \tag{9}$$

3. Repeat step 2 until convergence

For efficient way of solving (9) refer to [9]. In our experiments, we find that the neighborhood of each node can be estimated in a few seconds even when

the number of covariates is up to a thousand. A nice property of our algorithm is that the overall estimation procedure decouples to a collection of separate neighborhood estimation problems, which can be trivially parallelized. If we treat the neighborhood estimation as an atomic operation, the overall algorithm scales linearly as a product of the number of covariates p and the number of time points n, i.e. $\mathcal{O}(pn)$. For instance, the Drosophila data set in the application section contains 588 genes and 66 time points. The method `smooth` can estimate the neighborhood of one node, for all points in a regularization plane, in less than 1.5 hour.[1]

2.2 Structural changes in parameters

In this section, we give the estimation procedure of the non-zero pattern of $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$ under the assumption that the elements of $\boldsymbol{\theta}_u^t$ is a piecewise constant function, with pieces defined by the partition $\mathcal{B}_u$. Again, the estimation is performed node-wise and the estimators are combined using either Eq. (3) or Eq. (4). As opposed to the kernel smoothing estimator defined in Eq. (7), which gives the estimate at one time point τ, the procedure described below simultaneously estimates $\{\hat{\boldsymbol{\theta}}_u^t\}_{t\in\mathcal{T}_n}$. The estimators $\{\hat{\boldsymbol{\theta}}_u^t\}_{t\in\mathcal{T}_n}$ are defined as a minimizer of the following convex optimization objective:

$$\underset{\boldsymbol{\theta}_u^t\in\mathbb{R}^{p-1},\ t\in\mathcal{T}_n}{\operatorname{argmin}} \{\sum_{t\in\mathcal{T}_n}\gamma(\boldsymbol{\theta}_u^t;\mathbf{x}^t)+\lambda_1\sum_{t\in\mathcal{T}_n}||\boldsymbol{\theta}_u^t||_1+\lambda_{\text{TV}}\sum_{v\in V\setminus u}\text{TV}(\{\theta_{uv}^t\}_{t\in\mathcal{T}_n})\}, \quad (10)$$

where $\text{TV}(\{\theta_{uv}^t\}_{t\in\mathcal{T}_n}) := \sum_{i=2}^n |\theta_{uv}^{i/n} - \theta_{uv}^{(i-1)/n}|$ is the total variation penalty. We will refer to this approach of obtaining an estimator as `TV`. The penalty is structured as a combination of two terms. As mentioned before, the ℓ_1 norm of the parameters is used to regularize the solution towards estimators with lots of zeros and the regularization parameter λ_1 controls the number of non-zero elements. The second term penalizes the difference between parameters that are adjacent in time and, as a result, the estimated parameters have infrequent changes across time. This composite penalty, known as the "fused" Lasso penalty, was successfully applied in a slightly different setting of signal denoising (e.g., [25]) where it creates an estimate of the signal that is piecewise constant.

The optimization problem given in Eq. (10) is convex and can be solved using off-the-shelf interior point solver (e.g., the `CVX` package by [13]). However, for large scale problems (i.e., both p and n are large), interior point method can be computationally expensive, and we do not know of any specialized algorithm that can be used to solve (10) efficiently. Therefore, we propose a block-coordinate descent procedure which is much more efficient than the existing off-the-shelf solvers for large scale problems. Observe that the loss function can be decomposed as $\mathcal{L}(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}) = f_1(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}) + \sum_{v\in V\setminus u} f_2(\{\theta_{uv}^t\}_{t\in\mathcal{T}_n})$ for a smooth differentiable convex function $f_1(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}) = \sum_{t\in\mathcal{T}_n}\gamma(\boldsymbol{\theta}_u^t;\mathbf{x}^t)$ and a

[1] We have used a server with dual core 2.6GHz processor and 2GB RAM.

convex function $f_2(\{\theta_{uv}^t\}_{t\in\mathcal{T}_n}) = \lambda_1 \sum_{t\in\mathcal{T}_n} |\theta_{uv}^t| + \lambda_{\text{TV}}\, \text{TV}(\{\theta_{uv}^t\}_{t\in\mathcal{T}_n})$. [29] established that the block-coordinate descent converges for loss functions with such structure. Based on this observation we propose the following algorithm:

1. Set initial values: $\hat{\boldsymbol{\theta}}_u^{t,0} \leftarrow \mathbf{0}, \quad \forall t \in \mathcal{T}_n$
2. For each $v \in V\backslash u$, set the current estimates $\{\hat{\theta}_{uv}^{t,iter+1}\}_{t\in\mathcal{T}_n}$ as a solution to the following optimization procedure:

$$\min_{\{\theta^t\in\mathbb{R}\}_{t\in\mathcal{T}_n}} \left\{ \begin{array}{c} \sum_{t\in\mathcal{T}_n} \gamma\left(\hat{\theta}_{u,1}^{t,iter+1}, \ldots, \hat{\theta}_{u,v-1}^{t,iter+1}, \theta^t, \hat{\theta}_{u,v+1}^{t,iter}, \ldots, \hat{\theta}_{u,p-1}^{t,iter}; \mathbf{x}^t\right) \\ +\lambda_1 \sum_{t\in\mathcal{T}^n} |\theta^t| + \lambda_{\text{TV}}\, \text{TV}(\{\theta^t\}_{t\in\mathcal{T}_n}) \end{array} \right\} \tag{11}$$

3. Repeat step 2 until convergence

Using the proposed block-coordinate descent algorithm, we solve a sequence of optimization problems each with only n variables given in Eq. (11), instead of solving one big optimization problem with $n(n-1)$ variables given in Eq. (10). In our experiments, we find that the optimization in Eq. (10) can be estimated in an hour when the number of covariates is up to few hundreds and when the number of time points is also in hundreds. Here, the bottleneck is the number of time points. Observe that the dimensionality of the problem in Eq. (11) grows linearly with the number of time points. Again, the overall estimation procedure decouples to a collection of smaller problems which can be trivially parallelized. If we treat the optimization in Eq. (10) as an atomic operation, the overall algorithm scales linearly as a function of the number of covariates p, i.e. $\mathcal{O}(p)$. For instance, the Senate data set in the application section contains 100 Senators and 542 time points. It took about a day to solve the optimization problem in Eq. (10) for all points in the regularization plane.

2.3 Multiple observations

In the discussion so far, it is assumed that at any time point in $\mathcal{T}_n$ only one observation is available. There are situations with multiple observations at each time point, e.g., in a controlled repeated microarray experiment two samples obtained at a certain time point could be regarded as independent and identically distributed, and we discuss below how to incorporate such observations into our estimation procedures. Later, in Section 3 we empirically show how the estimation procedures benefit from additional observations at each time point.

For the estimation procedure given in Eq. (7) there are no modifications needed to accommodate multiple observations at a time point. Each additional sample will be assigned the same weight through the kernel function $K_h(\cdot)$. On the other hand, we need a small change in Eq. (10) to allow for multiple observations. The estimators $\{\hat{\boldsymbol{\theta}}_u^t\}_{t\in\mathcal{T}_n}$ are defined as follows:

$$\underset{\boldsymbol{\theta}_u^t\in\mathbb{R}^{p-1},\ t\in\mathcal{T}_n}{\text{argmin}} \{\sum_{t\in\mathcal{T}_n} \sum_{\mathbf{x}\in\mathcal{D}_n^t} \gamma(\boldsymbol{\theta}_u^t; \mathbf{x}) + \lambda_1 \sum_{t\in\mathcal{T}_n} ||\boldsymbol{\theta}_u^t||_1 + \lambda_{\text{TV}} \sum_{v\in V\backslash u} \text{TV}(\{\theta_{uv}^t\}_{t\in\mathcal{T}_n})\}, \tag{12}$$

where the set $\mathcal{D}_n^t$ denotes elements from the sample $\mathcal{D}_n$ observed at a time point t.

2.4 Choosing tuning parameters

Estimation procedures discussed in Section 2.1 and 2.2, `smooth` and `TV` respectively, require a choice of tuning parameters. These tuning parameters control sparsity of estimated graphs and the way the graph structure changes over time. The tuning parameter λ_1, for both `smooth` and `TV`, controls the sparsity of the graph structure. Large values of the parameter λ_1 result in estimates with lots of zeros, corresponding to sparse graphs, while small values result in dense models. Dense models will have a higher pseudo-likelihood score, but will also have more degrees of freedom. A good choice of the tuning parameters is essential in obtaining a good estimator that does not overfit the data, and balances between the pseudo-likelihood and the degrees of freedom. The bandwidth parameter h and the penalty parameter λ_{TV} control how similar are estimated networks that are close in time. Intuitively, the bandwidth parameter controls the size of a window around time point τ from which observations are used to estimate the graph G^τ. Small values of the bandwidth result in estimates that change often with time, while large values produce estimates that are almost time invariant. The penalty parameter λ_{TV} biases the estimates $\{\hat{\boldsymbol{\theta}}_u^t\}_{t\in\mathcal{T}_n}$ that are close in time to have similar values; large values of the penalty result in graphs whose structure changes slowly, while small values allow for more changes in estimates.

First, we discuss how to choose the penalty parameters λ_1 and λ_{TV} for the method `TV`. Observe that $\gamma(\boldsymbol{\theta}_u^t; \mathbf{x}^t)$ represents a logistic regression loss function when regressing a node u onto the other nodes $V\backslash u$. Hence, problems defined in Eq. (7) and Eq. (10) can be regarded as *supervised* classification problems, for which a number of techniques can be used to select the tuning parameters, e.g., cross-validation or held-out datasets can be used when enough data is available, otherwise, the BIC score can be employed. In this paper, we focus on the BIC score defined for $\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}$ as:

$$\text{BIC}(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}) := \sum_{t\in\mathcal{T}_n} \gamma(\boldsymbol{\theta}_u^t; \mathbf{x}^t) - \frac{\log n}{2}\text{Dim}(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}), \tag{13}$$

where $\text{Dim}(\cdot)$ denotes the degrees of freedom of the estimated model. Similar to [28], we adopt the following approximation to the degrees of freedom:

$$\text{Dim}(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}) = \sum_{t\in\mathcal{T}_n}\sum_{v\in V\backslash u} \mathbb{I}\left[\text{sign}(\theta_{uv}^t) \neq \text{sign}(\theta_{uv}^{t-1})\right] \times \mathbb{I}\left[\text{sign}(\theta_{uv}^t) \neq 0\right], \tag{14}$$

which counts the number of blocks on which the parameters are constant and not equal to zero. In practice, we average the BIC scores from all nodes and choose models according to the average.

Next, we address the way to choose the bandwidth h and the penalty parameter λ_1 for the method `smooth`. As mentioned earlier, the tuning of bandwidth

parameter h should trade off the smoothness of the network changes and the coverage of samples used to estimate the network. Using a wider bandwidth parameter provides more samples to estimate the network, but this risks missing sharper changes in the network; using a narrower bandwidth parameter makes the estimate more sensitive to sharper changes, but this also makes the estimate subject to larger variance due to the reduced effective sample size. In this paper, we adopt a heuristic for tuning the inital scale of the bandwidth parameter: we set it to be the median of the distance between pairs of time points. That is, we first form a matrix (d_{ij}) with its entries $d_{ij} := (t_i - t_j)^2$ $(t_i, t_j \in \mathcal{T}_n)$. Then the scale of the bandwidth parameter is set to the median of the entries in (d_{ij}). In our later simulation experiments, we find that this heuristic provides a good initial guess for h, and it is quite close to the value obtained via exhaustive grid search. For the method `smooth`, the BIC score for $\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}$ is defined as:

$$\mathrm{BIC}(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}) := \sum_{\tau\in\mathcal{T}_n}\sum_{t\in\mathcal{T}_n} w_t^\tau \gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t) - \frac{\log n}{2}\mathrm{Dim}(\{\boldsymbol{\theta}_u^t\}_{t\in\mathcal{T}_n}), \tag{15}$$

where $\mathrm{Dim}(\cdot)$ is defined in Eq. (14).

3 Simulation studies

We have conducted a small empirical study of the performance of methods `smooth` and `TV`. Our idea was to choose parameter vectors $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$, generate data according to the model in Eq. (1) using Gibbs sampling and try to recover the non-zero pattern of $\boldsymbol{\theta}^t$ for each $t \in \mathcal{T}_n$. Parameters $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$ are considered to be evaluations of the function $\boldsymbol{\theta}^t$ at $\mathcal{T}_n$ and we study two scenarios, as discussed in Section 2: $\boldsymbol{\theta}^t$ is a smooth function, $\boldsymbol{\theta}^t$ is a piecewise constant function. In addition to the methods `smooth` and `TV`, we will use the method of [23] to estimate a time-invariant graph structure, which we refer to as `static`. All the three methods estimate the graph based on node-wise neighborhood estimation, which, as discussed in Section 2, may produce asymmetric estimates. Solutions combined with the min operation in Eq. (3) are denoted as `****.MIN`, while those combined with the max operation in Eq. (4) are denoted as `****.MAX`.

We took the number of nodes $p = 20$, the maximum node degree $s = 4$, the number of edges $e = 25$ and the sample size $n = 500$. The parameter vectors $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$ and observation sequences are generated as follows:

1. Generate a random graph $\tilde{G}^0$ with 20 nodes and 15 edges: edges are added, one at a time, between random pairs of nodes that have the node degree less than 4. Next, randomly add 10 edges and remove 10 edges from $\tilde{G}^0$, taking care that the maximum node degree is still 4, to obtain $\tilde{G}^1$. Repeat the process of adding and removing edges from $\tilde{G}^1$ to obtain $\tilde{G}^2, \ldots, \tilde{G}^5$. We refer to these 6 graphs as the anchor graphs. We will randomly generate the prototype parameter vectors $\tilde{\boldsymbol{\theta}}^0, \ldots, \tilde{\boldsymbol{\theta}}^5$, corresponding to the anchor graphs, and then interpolate between them to obtain the parameters $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$.

2. Generate a prototype parameter vector $\tilde{\boldsymbol{\theta}}^i$ for each anchor graph $\tilde{G}^i$, $i \in \{0, \ldots, 5\}$, by sampling non-zero elements of the vector independently from $\mathsf{Unif}([0.5, 1])$. Then generate $\{\boldsymbol{\theta}^t\}_{t \in \mathcal{T}_n}$ according to one of the following two cases:

   - Smooth function: The parameters $\{\boldsymbol{\theta}^t\}_{t \in ((i-1)/5, i/5] \cap \mathcal{T}_n}$ are obtained by linearly interpolating 100 points between $\tilde{\boldsymbol{\theta}}^{i-1}$ and $\tilde{\boldsymbol{\theta}}^i$, $i \in \{1, \ldots, 5\}$.
   - Piecewise constant function: The parameters $\{\boldsymbol{\theta}^t\}_{t \in ((i-1)/5, i/5] \cap \mathcal{T}_n}$ are set to be equal to $(\tilde{\boldsymbol{\theta}}^{i-1} + \tilde{\boldsymbol{\theta}}^i)/2$, $i \in \{1, \ldots, 5\}$.

   Observe that after interpolating between the prototype parameters, a graph corresponding to $\boldsymbol{\theta}^t$ has 25 edges and the maximum node degree is 4.

3. Generate 10 independent samples at each $t \in \mathcal{T}_n$ according to $\mathbb{P}_{\boldsymbol{\theta}^t}$, given in Eq. (1), using Gibbs sampling.

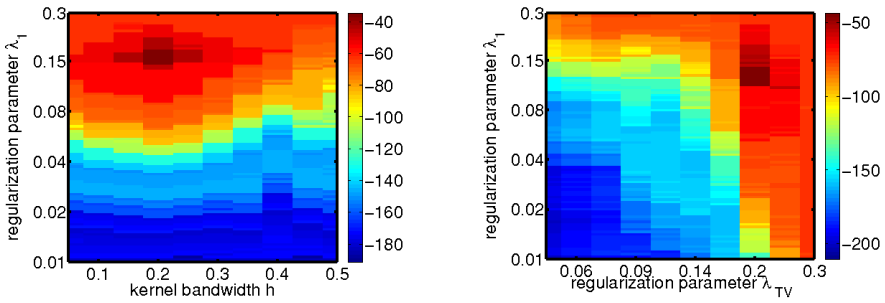
We estimate $\hat{G}^t$ for each $t \in \mathcal{T}_n$ with our `smooth` and `TV` methods, using $k \in \{1, \ldots, 10\}$ samples at each time point. The results are expressed in terms of the precision (Pre) and the recall (Rec) and $F1$ score, which is the harmonic mean of precision and recall, i.e., $F1 := 2 * \mathsf{Pre} * \mathsf{Rec}/(\mathsf{Pre} + \mathsf{Rec})$. Let $\hat{E}^t$ denote the estimated edge set of $\hat{G}^t$, then the precision is calculated as $\mathsf{Pre} := 1/n \sum_{t \in \mathcal{T}_n} |\hat{E}^t \cap E^t|/|\hat{E}^t|$ and the recall as $\mathsf{Rec} := 1/n \sum_{t \in \mathcal{T}_n} |\hat{E}^t \cap E^t|/|E^t|$. Furthermore, we report results averaged over 20 independent runs.

The tuning parameters h and λ_1 for `smooth`, and λ_1 and λ_{TV} for `TV` are chosen by maximizing the average BIC score,

$$\mathrm{BIC}_{\mathrm{avg}} := 1/p \sum_{u \in V} \mathrm{BIC}(\{\boldsymbol{\theta}_u^t\}_{t \in \mathcal{T}_n}),$$

over a grid of parameters. The bandwidth parameter h is searched over $\{0.05, 0.1, \ldots, 0.45, 0.5\}$ and the penalty parameter λ_{TV} over 10 points, equidistant on the log-scale, from the interval $[0.05, 0.3]$. The penalty parameter is searched over 100 points, equidistant on the log-scale, from the interval $[0.01, 0.3]$ for both `smooth` and `TV`. The same range is used to select the penalty parameter λ for the method `static` that estimates a time-invariant network. In our experiments, we use the Epanechnikov kernel $K(z) = 3/4 * (1 - z^2)\, \mathbb{1}\{|z| \leq 1\}$ and we remind our reader that $K_h(\cdot) = K(\cdot/h)$. For illustrative purposes, in Figure 1 we plot the $\mathrm{BIC}_{\mathrm{avg}}$ score over the grid of tuning parameters.

First, we discuss the estimation results when the underlying parameter vector changes smoothly. See Figure 2 for results. It can be seen that as the number of the *i.i.d.* observations at each time point increases, the performance of both methods `smooth` and `TV` increases. On the other hand, the performance of the method `static` does not benefit from additional *i.i.d.* observations. This observation should not be surprising as the time-varying network models better fit the data generating process. When the underlying parameter vector $\boldsymbol{\theta}^t$ is a smooth function of time we expect that the method `smooth` would have



(a) Average BIC score

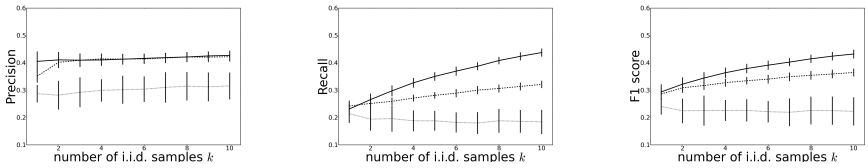
(b) Average BIC score

Figure 1: Plot of the BIC_{avg} score over the regularization plane. The parameter vector θ^t is a smooth function of time and at each time point there is one observation. (a) The graph structure recovered using the method smooth. (b) The graph structure recovered using the method TV.

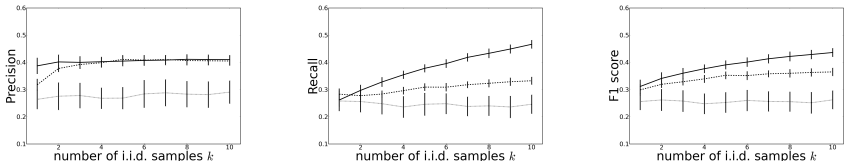
a faster convergence and better performance, which can be seen in Figure 2. There are some differences between the estimates obtained through `MIN` and `MAX` symmetrization. In our limited numerical experience, we have seen that `MAX` symmetrization outperforms `MIN` symmetrization. `MIN` symmetrization is more conservative in including edges to the graph and seems to be more susceptible to noise.

Next, we discuss the estimation results when then the underlying parameter vector is piecewise constant function. See Figure 3 for results. Again, both performance of the method `smooth` and of the method `TV` improve as there are more independent samples at different time points, as opposed to the method `static`. It is worth noting that the empirical performance of `smooth` and `TV` is very similar in the setting when $\boldsymbol{\theta}^t$ is a piecewise constant function of time, with the method `TV` performing marginally better. This may be a consequence of the way we present results, averaged over all time points in $\mathcal{T}_n$. A closer inspection of the estimated graphs shows that the method `smooth` poorly estimates graph structure close to time point at which the parameter vector changes abruptly (results not shown).

We have decided to perform simulation studies on Erdös-Rényi graphs, while real-world graphs are likely to have different properties, such as a scale-free network with a long tail in its degree distribution. From a theoretical perspective (see Section 5), our method can still recover the true structure of these networks regardless of the degree distribution, although for a more complicated model, we may need more samples in order to achieve this. [22] proposed a joint sparse regression model, which performs better than the neighborhood selection method when estimating networks with hubs (nodes with very high degree) and scale-free networks. For such networks, we can extended their model to our time-varying setting, and potentially make more efficient use of the samples,

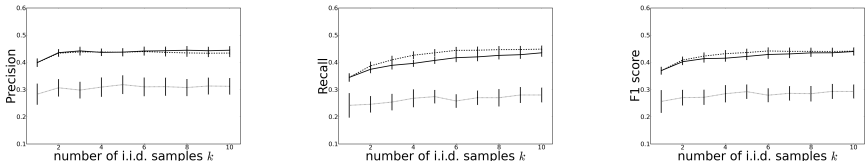


(a) Precision (b) Recall (c) F1 score

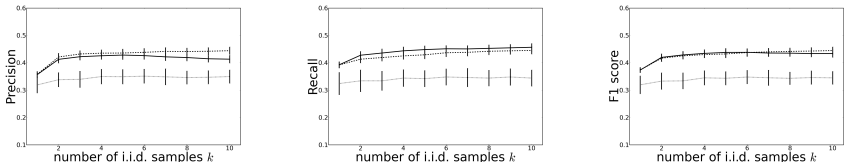


(d) Precision (e) Recall (f) F1 score

Figure 2: Results of estimation when the underlying parameter $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$ changes smoothly with time. The upper row consists of results when the graph is estimated combining the neighborhoods using the min operation, while the lower row consists of results when the max operation is used to combine neighborhoods. Precision, recall and F1 score are plotted as the number of *i.i.d.* samples k at each time point increases from 1 to 10. The solid, dashed, and dotted lines denote results for smooth, TV, and static, respectively.



(a) Precision (b) Recall (c) F1 score



(d) Precision (e) Recall (f) F1 score

Figure 3: Results of estimation when the underlying parameter $\{\boldsymbol{\theta}^t\}_{t\in\mathcal{T}_n}$ is a piecewise constant function of time. The upper row consists of results when the graph is estimated combining the neighborhoods using the min operation, while the lower row consists of results when the max operation is used to combine neighborhoods. Precision, recall and F1 score are plotted as the number of *i.i.d.* samples k at each time point increases from 1 to 10. The solid, dashed, and dotted lines denote results for smooth, TV, and static, respectively.

however, we do not pursue this direction here.

4 Applications to real data

In this section we present the analysis of two real data sets using the algorithms presented in Section 2. First, we present the analysis of the senate data consisting of Senators' votes on bills during the 109th Congress. The second data set consists of expression levels of more than 4000 genes from the life cycle of *Drosophila melanogaster*.

4.1 Senate Voting Records Data

The US senate data consists of voting records from 109th congress (2005 - 2006)[2]. There are 100 senators whose votes were recorded on the 542 bills. Each senator corresponds to a variable, while the votes are samples recorded as -1 for no and 1 for yes. This data set was analyzed in [2], where a static network was estimated. Here, we analyze this data set in a time varying framework in order to discover how the relationship between senators changes over time.

This data set has many missing values, corresponding to votes that were not cast. We follow the approach of [2] and fill those missing values with (-1). Bills were mapped onto the $[0, 1]$ interval, with 0 representing Jan 1st, 2005 and 1 representing Dec 31st, 2006. We use the Epanechnikov kernel for the method `smooth`. The tuning parameters are chosen optimizing the average BIC score over the same range as used for the simulations in Section 3. For the method `smooth`, the bandwidth parameter was selected as $h = 0.174$ and the penalty parameter $\lambda_1 = 0.195$, while penalty parameters $\lambda_1 = 0.24$ and $\lambda_{\text{TV}} = 0.28$ were selected for the method `TV`. In the figures in this section, we use pink square nodes to represent republican Senators and blue circle nodes to represent democrat Senators.

A first question is whether the learned network reflect the political division between Republicans and Democrats. Indeed, at any time point t, the estimated network contains few clusters of nodes. These clusters consist of either Republicans or Democrats connected to each others, see Figure 4. Furthermore there are very few links connecting different clusters. We observe that most Senators vote similarly to other members of their party. Links connecting different clusters usually go through senators that are members of one party, but have views more similar to the other party, e.g. Senator Ben Nelson or Senator Chafee. Note that we do not necessarily need to estimate a time evolving network to discover this pattern of political division, as they can also be observed from a time-invariant network, e.g. see [2].

Therefore, what is more interesting is whether there is any time evolving pattern. To show this, we examine neighborhoods of Senators Jon Corzine and Bob Menendez. Senator Corzine stepped down from the Senate at the end of the 1st Session in the 109th Congress to become the Governor of New Jersey.

[2]The data can be obtain from the U.S. Senate web page `http://www.senate.gov`

His place in the Senate was filled by Senator Menendez. This dynamic change of interactions can be well captured by the time-varying network (Figure 5). Interestingly, we can see that Senator Lautenberg who used to interact with Senator Corzine switch to Senator Menendez in response to this event.

Another interesting question is whether we can discover senators with swaying political stance based on time evolving networks. We discover that Senator Ben Nelson and Lincoln Chafee fall into this category. Although, Senator Ben Nelson is a Democrat from Nebraska, he is considered as one of the most conservative Democrats in the Senate. Figure 6 presents neighbors at distance two or less of Senator Ben Nelson at two time points, one during the 1st Session and one during the 2nd Session. As a conservative Democrat, he is connected to both Democrats and Republicans since he shares views with both parties. This observation is supported by Figure 6(a) which presents his neighbors during the 1st Session. It is also interesting to note that during the second session, his views drifted more towards the Republicans (Figure 6(b)). For instance, he voted against abortion and withdrawal of most combat troops from Iraq, which are both Republican views.

In contrast, although Senator Lincoln Chafee is a Republican, his political view grew increasingly Democratic. Figure 7 presents neighbors of Senator Chafee at three time points during the 109th Congress. We observe that his neighborhood includes an increasing amount of Democrats as time progresses during the 109th Congress. Actually, Senator Chafee later left the Republican Party and became an independent in 2007. Also, his view on abortion, gay rights and environmental policies are strongly aligned with those of Democrats, which is also consistently reflected in the estimated network. We emphasize that these patterns about Senator Nelson and Chafee could not be observed in a static network.

4.2 Gene Regulatory Networks of Drosophila Melanogaster

In this section, we used the kernel reweighting approach to reverse engineer the gene regulatory networks of *Drosophila melanogaster* from a time series of gene expression data measured during its full life cycle. Over the developmental course of *Drosophila melanogaster*, there exist multiple underlying "themes" that determine the functionalities of each gene and their relationships to each other, and such themes are dynamical and stochastic. As a result, the gene regulatory networks at each time point are context-dependent and can undergo systematic rewiring, rather than being invariant over time. In a seminal study by [20], it was shown that the "active regulatory paths" in the gene regulatory networks of *Saccharomyces cerevisiae* exhibit topological changes and hub transience during a temporal cellular process, or in response to diverse stimuli. We expect similar properties can also be observed for the gene regulatory networks of *Drosophila melanogaster*.

We used microarray gene expression measurements from [1] as our input data. In such an experiment, the expression levels of 4028 genes are simultaneously measured at various developmental stages. Particularly, 66 time points are

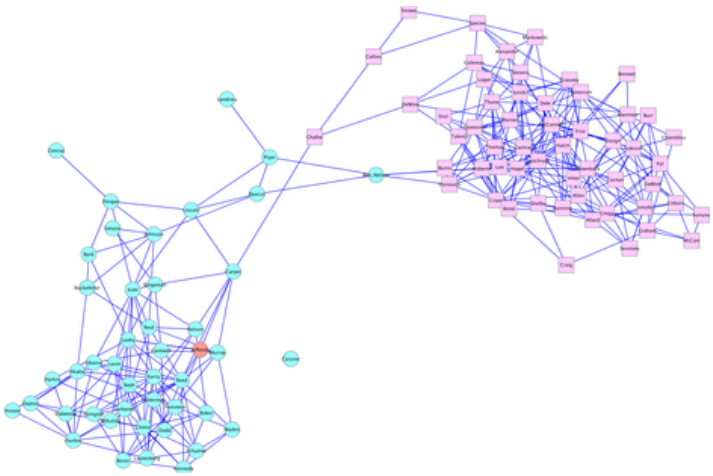
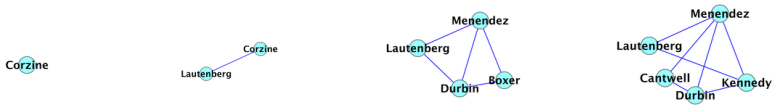
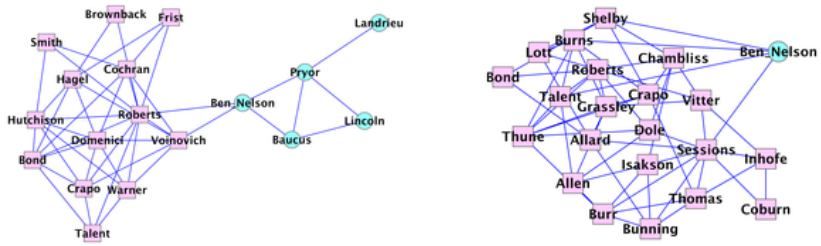


Figure 4: 109th Congress, Connections between Senators in April 2005. Democrats are represented with blue circles, Republicans with pink squares and the red circle represent independent Senator Jeffords.



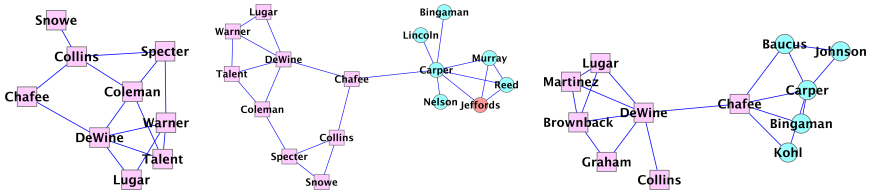
(a) March 2005 (b) August 2005 (c) March 2006 (d) August 2006

Figure 5: Direct neighbors of the node that represents Senator Corzine and Senator Menendez at four different time points. Senator Corzine stepped down at the end of the 1st Session and his place was taken by Senator Menendez, which is reflected in the graph structure.



(a) May 2005 (b) August 2006

Figure 6: Neighbors of Senator Ben Nelson (distance two or lower) at the beginning of 109th Congress and at the end of 109th Congress. Democrats are represented with blue circles, Republicans with pink squares. The estimated neighborhood in August 2006 consists only of Republicans, which may be due to the type of bills passed around that time on which Senator Ben Nelson had similar view as other Republicans.



(a) February 2005 (b) October 2005 (c) July 2006

Figure 7: Neighbors of Senator Chafee (distance two or lower) at different time points during 109th Congress. Democrats are represented with blue circles, Republicans with pink squares and the red circle represent independent Senator Jeffords.

chosen during the full developmental cycle of *Drosophila melanogaster*, spanning across four different stages, *i.e.* embryonic (1–30 time point), larval (31–40 time point), pupal (41–58 time points) and adult stages (59–66 time points). In this study, we focused on 588 genes that are known to be related to developmental process based on their gene ontologies.

Usually, the samples prepared for microarray experiments are a mixture of tissues with possibly different expression levels. This means that microarray experiments only provide rough estimates of the average expression levels of the mixture. Other sources of noise can also be introduced into the microarray measurements during, for instance, the stage of hybridization and digitization. Therefore, microarray measurements are far from the exact values of the expression levels, and it will be more robust if we only consider the binary state of the the gene expression: either being up-regulated or down-regulated. For this reason, we binarize the gene expression levels into $\{-1, 1\}$ (-1 for down-regulated and 1 for up-regulated). We learned a sequence of binary MRFs from these time series.

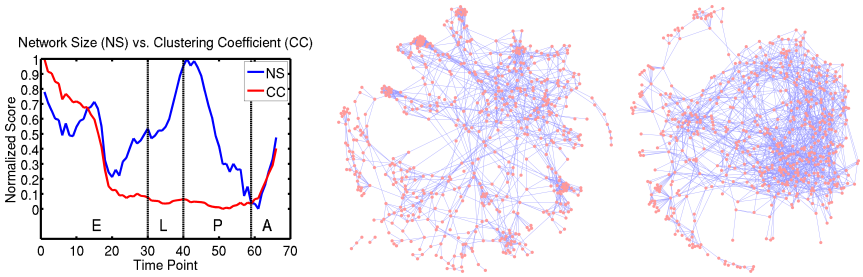
First, we study the global pattern of the time evolving regulatory networks. In Figure 8(a), we plotted two different statistics of the reversed engineered gene regulatory networks as a function of the developmental time point (1–66). The first statistic is the network size as measured by the number of edges; and the second is the average local clustering coefficient as defined by [32]. For comparison, we normalized both statistics to the range between $[0, 1]$. It can be seen that the network size and its local clustering coefficient follow very different trajectories during the developmental cycle. The network size exhibits a wave structure featuring two peaks at mid-embryonic stage and the beginning of pupal stage. Similar pattern of gene activity has also been observed by [1]. In contrast, the clustering coefficients of the dynamic networks drop sharply after the mid-embryonic stage, and they stay low until the start of the adult stage. One explanation is that at the beginning of the development process, genes have a more fixed and localized function, and they mainly interact with other genes with similar functions; however, after mid-embryonic stage, genes become more versatile and involved in more diverse roles to serve the need of rapid development; as the organism turns into an adult, its growth slows down and each gene is restored to its more specialized role. To illustrate how the network properties change over time, we visualized two networks from mid-embryonic stage (time point 15) and mid-pupal stage (time point 45) using spring layout algorithm in Figure 8(b) and 8(c) respectively. Although the size of the two networks are comparable, tight local clusters of interacting genes are more visible during mid-embryonic stage than mid-pupal stage, which is consistent with the evolution local clustering coefficient in Figure 8(a).

To judge whether the learned networks make sense biologically, we zoom into three groups of genes functionally related to different stages of development process. In particular, the first group (30 genes) is related to embryonic development based on their functional ontologies; the second group (27 genes) is related to post-embryonic development; and the third group (25 genes) is related to muscle development. For each group, we use the number of within

group connections plus all its outgoing connections to describe the activitiy of each group of genes (for short, we call it interactivity). In Figure 9, we plotted the time courses of interactivity for the three groups respectively. For comparison, we normalize all scores to the range of $[0, 1]$. We see that the time courses have a nice correspondence with their supposed roles. For instance, embryonic development genes have the highest interactivity during embryonic stage, and post-embryonic genes increase their interacativity during larval and pupal stage. The muscle development genes are less specific to certain developmental stages, since they are needed across the developmental cycle. However, we see its increased activity when the organism approaches its adult stage where muscle development becomes increasingly important.

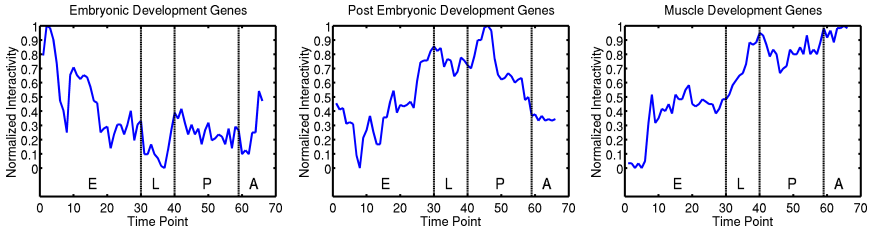
The estimated networks also recover many known interactions between genes. In recovering these known interactions, the dynamic networks also provide additional information as to when interactions occur during development. In Figure 10, we listed these recovered known interactions and the precise time when they occur. This also provides a way to check whether the learned networks are biologically plausible given the prior knowledge of the actual occurence of gene interactions. For instance, the interaction between genes msn and dock is related to the regulation of embryonic cell shape, correct targeting of photoreceptor axons. This is very consistent with the timeline provided by the dynamic networks. A second example is the interaction between genes sno and Dl which is related to the development of compound eyes of *Drosophila*. A third example is between genes caps and Chi which are related to wing development during pupal stage. What is most interesting is that the dynamic networks provide timelines for many other gene interactions that have not yet been verified experimentally. This information will be a useful guide for future experiments.

We further studied the relations between 130 transcriptional factors (TF). The network contains several cluster of transcriptional cascades, and we will present the detail of the largest transcriptional factor cascade involving 36 transcriptional factors (Figure 11). This cascade of TFs is functionally very coherent, and many TFs in this network play important roles in the nervous system and eye development. For example, Zn finger homeodomain 1 (zhf1), brinker (brk), charlatan (chn), decapentaplegic (dpp), invected (inv), forkhead box, subgroup 0 (foxo), Optix, eagle (eg), prospero (pros), pointed (pnt), thickveins (tkv), extra macrochaetae (emc), lilliputian (lilli), doublesex (dsx) are all involved in nervous and eye development. Besides functional coherence, the networks also reveals the dynamic nature of gene regulation: some relations are persistent across the full developmental cycle while many others are transient and specific to certain stages of development. For instance, five transcriptional factors, brk-pnt-zfh1-pros-dpp, form a long cascade of regulatory relations which are active across the full developmental cycle. Another example is gene Optix which are active across the full developmental cycle and serves as a hub for many other regulatory relations. As for transience of the regulatory relations, TFs to the right of Optix hub reduced in their activity as development proceeds to later stage. Furthermore, Optix connects two disjoint cascade of gene regulations to



(a) Network statistics (b) Mid-embryonic stage (c) Mid-pupal stage

Figure 8: Characteristic of the dynamic networks estimated for the genes related to developmental process. (a) Plot of two network statistics as functions of development time line. Network size ranges between 1712 and 2061 over time, while local clustering coefficient ranges between 0.23 and 0.53 over time; To focus on relative activity over time, both statistics are normalized to the range between 0 and 1. (b) and (c) visualization of two example of networks from different time point. We can see that network size can evolve in a very different way from the local clustering coefficient.



(a) (b) (c)

Figure 9: Interactivity of 3 groups of genes related to (a) embryonic development (ranging between 169 and 241); (b) post-embryonic development (ranging between 120 and 210) and (c) muscle development (ranging between 29 and 89). To focus on the relative activity over time, we normalize the score to $[0, 1]$. The higher the interactivity, the more active the group of genes. The interactivities of these three groups are very consistent with their functional annotations.

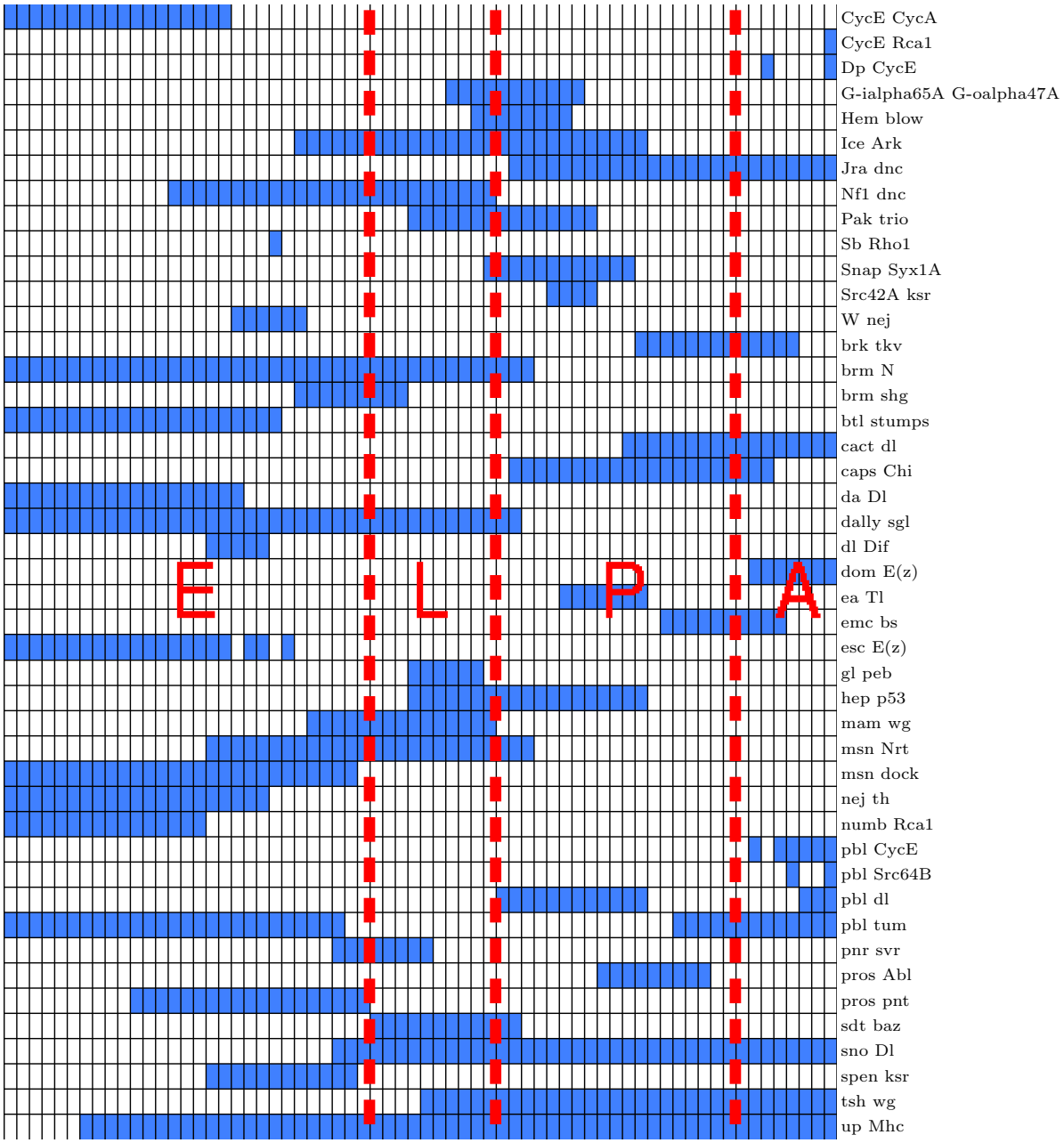
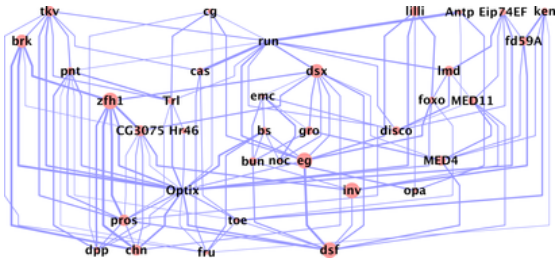
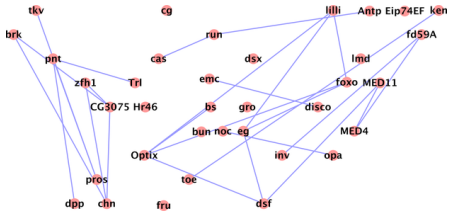


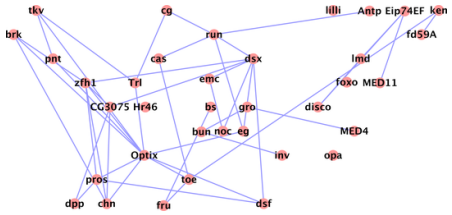
Figure 10: Timeline of 45 known gene interactions. Each cell in the plot corresponds to one gene pair of gene interaction at one specific time point. The cells in each row are ordered according to their time point, ranging from embryonic stage (E) to larval stage (L), to pupal stage (P), and to adult stage (A). Cells colored blue indicate the corresponding interaction listed in the right column is present in the estimated network; blank color indicates the interaction is absent.



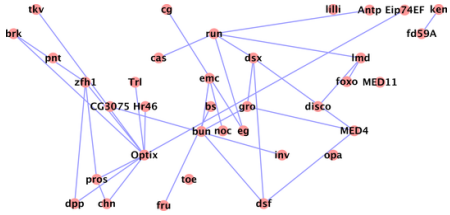
(a) Summary network



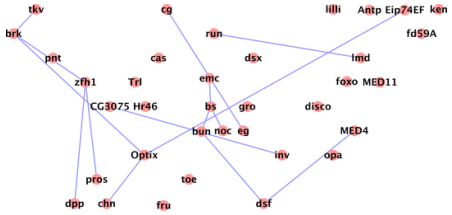
(b) Time point 15 (mid-embryonic stage)



(c) Time point 35 (mid-larval stage)



(d) Time point 49 (mid-pupal stage)



(e) Time point 62 (mid-adult stage)

Figure 11: The largest transcriptional factors (TF) cascade involving 36 transcriptional factors. (a) The summary network is obtained by summing the networks from all time points. Each node in the network represents a transcriptional factor, and each edge represents an interaction between them. On different stages of the development, the networks are different, (b,c,d,e) shows representative networks for the embryonic, larval, pupal and adult stage of the development respectively.

its left and right side after embryonic stage.

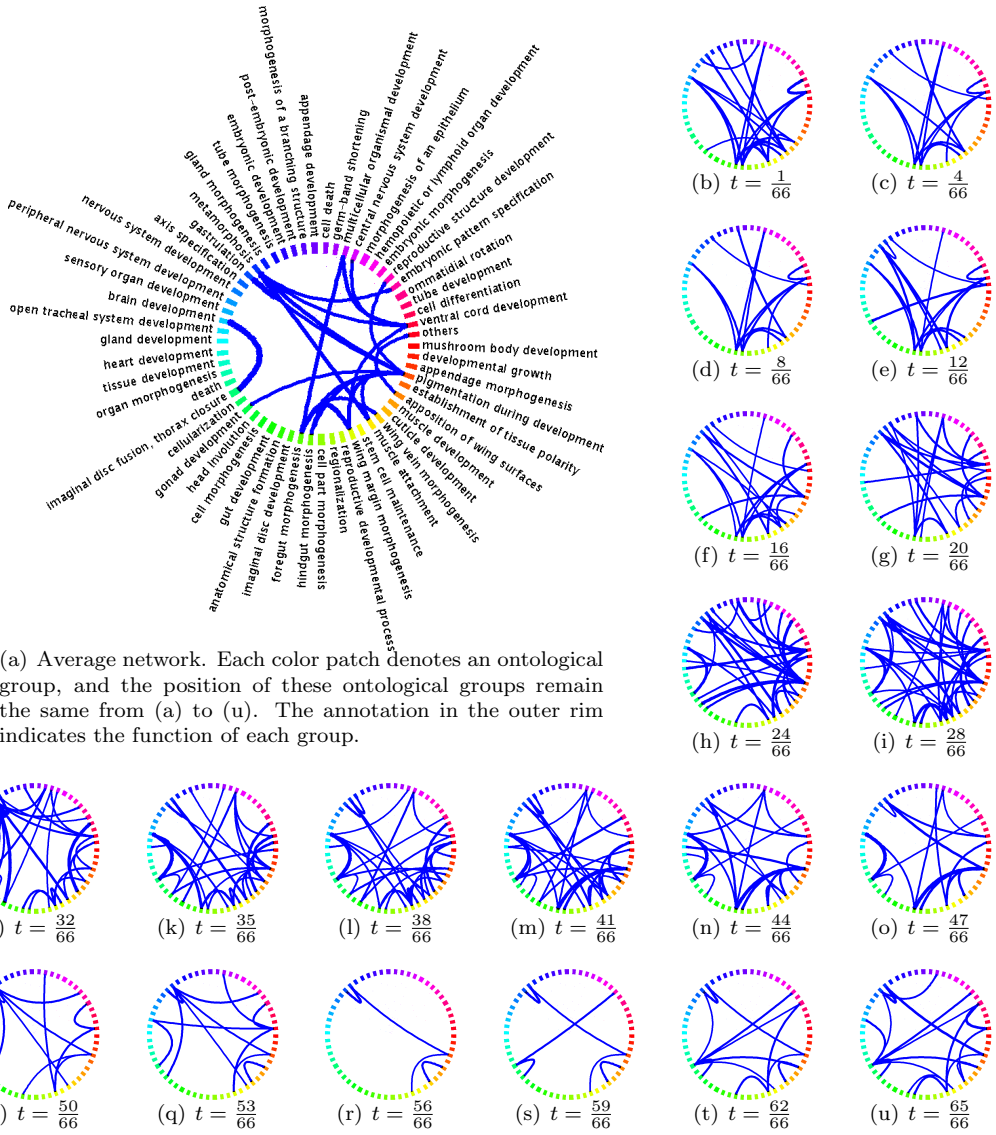
The dynamic networks also provide an overview of the interactions between genes from different functional groups. In Figure 12, we grouped genes according to 58 ontologies and visualized the connectivity between groups. We can see that large topological changes and network rewiring occur between functional groups. Besides expected interactions, the figure also reveals many seemingly unexpected interactions. For instance, during the transition from pupa stage to adult stage, *Drosophila* is undergoing a huge metamorphosis. One major feature of this metamorphosis is the development of the wing. As can be seen from Figure 12(r) and 12(s), genes related to metamorphosis, wing margin morphogenesis, wing vein morphogenesis and apposition of wing surfaces are among the most active group of genes, and they carry their activity into adult stage. Actually, many of these genes are also very active during early embryonic stage (for example, Figure 12(b) and 12(c)); the difference is though they interact with different groups of genes. On one hand, the abundance of the transcripts from these genes at embryonic stage is likely due to maternal deposit [1]; on the other hand, this can also be due to the diverse functionalities of these genes. For instance, two genes related to wing development, held out wings (how) and tolloid (td), also play roles in embryonic development.

5 Some properties of the algorithms

In this section we discuss some theoretical guarantees of the proposed algorithms. The most challenging aspect in estimating time-varying graphs is that the dimension of the data p can be much larger than the size of the sample n ($p \gg n$), and there is usually only one sample per time point. For example, in a genome-wide reverse engineering task, the number of genes can be well over ten thousand ($p > 10,000$), while the total number of microarray measurements is only in hundreds ($n \sim 100$) and the measurements are collected at different developmental stages. Then, the question is, what are the sufficient conditions under which our algorithms recovers the sequence of unknown graphs $\{G^t\}_{t\in\mathcal{T}_n}$ correctly.

5.1 Recovery under smooth changes

For convenience, we restate the estimation problem from Section 2. Recall that $\mathcal{D}_n = \{\mathbf{x}^t \sim \mathbb{P}_{\boldsymbol{\theta}^t} | t \in \mathcal{T}_n\}$ denotes a sample of n data points, which are sampled independently from $\mathbb{P}_{\boldsymbol{\theta}^t}$ at discrete time steps indexed by $\mathcal{T}_n$, where $\mathbb{P}_{\boldsymbol{\theta}^t}$ is given in Eq. (1). The problem of the graph structure estimation associated with the distribution $\mathbb{P}_{\boldsymbol{\theta}^\tau}$, at any given time point $\tau \in [0,1]$, is cast as the problem of estimating the non-zero pattern of the vector $\boldsymbol{\theta}^\tau$, i.e., locations of non-zero elements of $\boldsymbol{\theta}^\tau$. A stronger notion of structure estimation is that of *signed edge recovery*; for a given graphical model G^τ with parameter $\boldsymbol{\theta}^\tau$, we define the signed



(a) Average network. Each color patch denotes an ontological group, and the position of these ontological groups remain the same from (a) to (u). The annotation in the outer rim indicates the function of each group.

(b) $t = \frac{1}{66}$ (c) $t = \frac{4}{66}$ (d) $t = \frac{8}{66}$ (e) $t = \frac{12}{66}$ (f) $t = \frac{16}{66}$ (g) $t = \frac{20}{66}$ (h) $t = \frac{24}{66}$ (i) $t = \frac{28}{66}$ (j) $t = \frac{32}{66}$ (k) $t = \frac{35}{66}$ (l) $t = \frac{38}{66}$ (m) $t = \frac{41}{66}$ (n) $t = \frac{44}{66}$ (o) $t = \frac{47}{66}$ (p) $t = \frac{50}{66}$ (q) $t = \frac{53}{66}$ (r) $t = \frac{56}{66}$ (s) $t = \frac{59}{66}$ (t) $t = \frac{62}{66}$ (u) $t = \frac{65}{66}$

Figure 12: Interactions between gene ontological groups related to developmental process undergo dynamic rewiring. The weight of an edge between two ontological groups is the total number of connection between genes in the two groups. In the visualization, the width of an edge is proportional to its edge weight. We thresholded the edge weight at 30 in (b)-(u) so that only those interactions exceeding this number are displayed. The average network in (a) is produced by averaging the networks underlying (b)-(u). In this case, the threshold is set to 20 instead.

edge vector $SE^\tau \in \mathbb{R}^{\binom{p}{2}}$ as:

$$SE^\tau = \begin{cases} \text{sign}(\theta^\tau_{uv}) & \text{if } (u,v) \in E^\tau \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

In Section 2 we discussed the problem of graph recovery as recovering the vector $|SE^\tau|$ of absolute values. The guarantees we give here are stronger and address estimation of the signed edge vector.

Let $\tau \in [0,1]$ be any given time point for which we are interested in estimating the structure of the graph G^τ from the sample $\mathcal{D}_n$. Typically, the structure of the graph is estimated for every $\tau \in \mathcal{T}_n$. Observe the signed edge vector SE^τ of a graph G^τ can be recovered from the set of neighboring edges $S^\tau(u) = \{(u,v) \ : \ (u,v) \in E^\tau\}$ $(u \in V)$ and the correct signs $\text{sign}(\theta^\tau_{uv})$ for all $(u,v) \in S^\tau(u)$. We define the set of *signed neighboring edges* as

$$S^\tau_\pm(u) := \{(\text{sign}(\theta^\tau_{uv}), (u,v)) \ : \ (u,v) \in S^\tau(u)\}.$$

The set of signed neighboring edges $S^\tau_\pm(u)$ can be determined from the signs of elements of the $(p-1)$-dimensional subvector of parameters

$$\boldsymbol{\theta}^\tau_u := \{\theta^\tau_{uv} \ : \ v \in V \backslash u\}$$

associated with vertex u. Under the model (1), the conditional distribution of X^τ_u given other variables $\mathbf{X}^\tau_{\backslash u} := \{X^\tau_v \ : \ v \in V \backslash u\}$ is given in Eq. (5) and the log-likelihood, for one data-point $t \in \mathcal{T}_n$, has the form given in Eq. (6). For an arbitrary point of interest $\tau \in [0,1]$, this log-likelihood suggest an estimator $\hat{\boldsymbol{\theta}}^\tau_u$ of the sign-pattern of the vector $\boldsymbol{\theta}^\tau_u$ as the solution to the following convex program (already given in Eq. (7)):

$$\hat{\boldsymbol{\theta}}^\tau_u = \min_{\boldsymbol{\theta}_u \in \mathbb{R}^{p-1}} \{\ell(\boldsymbol{\theta}_u; \mathcal{D}_n) + \lambda_n ||\boldsymbol{\theta}_u||_1\} \tag{17}$$

where

$$\ell(\boldsymbol{\theta}_u; \mathcal{D}_n) = -\sum_{t \in \mathcal{T}_n} w^\tau_t \gamma(\boldsymbol{\theta}_u; \mathbf{x}^t) \tag{18}$$

and the weights are defined as

$$w^\tau_t = \frac{K_h(t-\tau)}{\sum_{t' \in \mathcal{T}_n} K_h(t'-\tau)}$$

for a symmetric nonnegative kernel $K_h(\cdot) = K(\cdot/h)$. Note that in the objective given in (17) we approximate the function $\boldsymbol{\theta}^t_u : \mathbb{R} \mapsto \mathbb{R}^{p-1}$ around the point τ with a constant $\hat{\boldsymbol{\theta}}^\tau_u \in \mathbb{R}^{p-1}$. The program (17) is convex, but not differentiable, because of the ℓ_1 norm. The minimum over $\boldsymbol{\theta}_u$ is always achieved, as the problem can be cast as a constrained optimization problem over the ball $||\boldsymbol{\theta}_u||_1 \leq C(\lambda_n)$ and the claim follows from the Weierstrass theorem.

Let $\hat{\boldsymbol{\theta}}^\tau_u$ be a minimizer of (17). The convex program (17) does not necessarily have a unique optimum, but as we will prove shortly, in the regime of interest

any two solutions will have non-zero elements in the same positions. Based on the vector $\hat{\boldsymbol{\theta}}_u^\tau$, we construct the estimate of the signed neighborhood:

$$\hat{S}_{\pm}^{\tau}(u) := \left\{ (\text{sign}(\hat{\theta}_{uv}^{\tau}), (u,v)) \ : \ v \in V \backslash u, \ \hat{\theta}_{uv}^{\tau} \neq 0 \right\}. \tag{19}$$

The structure of graph G^τ is consistently estimated if every signed neighborhood is recovered, i.e. $\hat{S}_{\pm}^{\tau}(u) = S_{\pm}^{\tau}(u)$ for all $u \in V$.

5.1.1 Main theoretical result

In this section, we describe under which conditions the graph structure can be recovered. We give the conditions under which the estimation procedure estimates the unknown graph structure consistently. Using notation from the last section, we give conditions under which the estimator $\widehat{SE}_n^\tau$ satisfies the following

$$\mathbb{P}[\widehat{SE}_n^\tau = SE^\tau] \to 1, \quad \text{as } n \to +\infty.$$

This property is known as *sparsistency*. We will mainly be interested in the high-dimensional case, where the dimension $p = p_n$ is comparable or even larger than the sample size n. It is of great interest to understand the performance of the estimator under this assumption, since in many real world scenarios the dimensionality of data is large. Our analysis is asymptotic and we consider the model dimension $p = p_n$ to grow at a certain rate as the sample size grows. This essentially allows us to consider more "complicated" models as we observe more data points. Another quantity that will describe the complexity of the model is the maximum node degree $s = s_n$, which is also considered as a function of the sample size. The main result describes the scaling of the triple (n, p_n, s_n) under which the estimation procedure given in the previous section estimates the graph structure consistently.

Since our main interest is in estimating the structure of a high-dimensional graph from a small size sample, we assume that the true structure of the graph is sparse, i.e., we assume that each node has a small number of adjacent edges. The ℓ_1 regularization procedures have been proved very successful as model selection techniques in a variety of problems, and, as we show here, our method is successful in estimating the time-varying graph structure.

In order to estimate the non-zero pattern of the vector $\boldsymbol{\theta}_u^\tau$ for each node $u \in V$ we need to impose regularity conditions on the covariates in the model (5). We express these conditions in terms of the Hessian of the log-likelihood function as evaluated at the true model parameter, i.e., the Fisher information matrix. The Fisher information matrix $\mathbf{Q}_u^\tau \in \mathbb{R}^{(p-1)\times(p-1)}$ is a matrix defined for each node $u \in V$ as:

$$\begin{aligned} \mathbf{Q}_u^\tau :&= \mathbb{E}[\nabla^2 \log \mathbb{P}_{\boldsymbol{\theta}_u^\tau}[X_u | \mathbf{X}_{\backslash u}]] \\ &= \mathbb{E}[\eta(\mathbf{X}; \boldsymbol{\theta}_u^\tau) \mathbf{X}_{\backslash u} \mathbf{X}'_{\backslash u}], \end{aligned} \tag{20}$$

where

$$\eta(\mathbf{x}; \boldsymbol{\theta}_u) := \frac{4 \exp(2 x_u \langle \boldsymbol{\theta}_u, \mathbf{x}_{\backslash u} \rangle)}{(\exp(2 x_u \langle \boldsymbol{\theta}_u, \mathbf{x}_{\backslash u} \rangle) + 1)^2}$$

is the variance function. We write $\mathbf{Q}^\tau := \mathbf{Q}^\tau_u$ and assume that the following assumptions hold for each node $u \in V$.

A1: Dependency condition There exist constants $C_{\min}, D_{\min}, D_{\max} > 0$ such that

$$\Lambda_{\min}(\mathbf{Q}^\tau_{SS}) \geq C_{\min}$$

and

$$\Lambda_{\min}\left(\mathbf{\Sigma}^\tau\right) \geq D_{\min}, \quad \Lambda_{\max}\left(\mathbf{\Sigma}^\tau\right) \leq D_{\max},$$

where $\mathbf{\Sigma}^\tau = \mathbb{E}_{\boldsymbol{\theta}^\tau}[\mathbf{X}^\tau \mathbf{X}^{\tau'}]$. Here $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalue of a matrix.

A2: Incoherence condition There exists an incoherence parameter $\alpha \in (0, 1]$ such that

$$|||\mathbf{Q}^\tau_{S^c S}(\mathbf{Q}^\tau_{SS})^{-1}|||_\infty \leq 1 - \alpha,$$

where, for a matrix $A \in \mathbb{R}^{a \times b}$, the ℓ_∞ matrix norm is defined as $|||A|||_\infty := \max_{i \in \{1,\ldots,a\}} \sum_{j=1}^{b} |a_{ij}|$.

With some abuse of notation, when defining assumptions A1 and A2, we use the index set $S := S^\tau(u)$ to denote nodes adjacent to the node u at time τ. For example, if $s = |S|$, then $\mathbf{Q}^\tau_{SS} \in \mathbb{R}^{s \times s}$ denotes the sub-matrix of $\mathbf{Q}^\tau$ indexed by S.

As in the structure estimation of the invariant MRF from an *i.i.d.* sample [23], the Fisher information matrix $\mathbf{Q}^\tau$, associated with the local conditional probability, plays very important role in determining success of the method. It can also be regarded as a counterpart of the covariance matrix $\mathbb{E}[\mathbf{X}^\tau \mathbf{X}^{\tau'}]$ of Gaussian graphical models. Note that the conditions A1 and A2 are symbolically the same as for the *i.i.d.* case, when the graph is invariant over time [23], with the difference that we assume that the conditions hold for the time point of interest τ at which we want to recover the graph structure. Condition A1 assures that the relevant features are not too correlated. Condition A2 assures that the irrelevant features do not have to strong effect onto the relevant features. Similar type of condition has been proposed also in the case of the linear regression (e.g. [21]).

Next, we assume that the distribution $\mathbb{P}_{\boldsymbol{\theta}^t}$ changes smoothly over time, which we express in the following form, for every node $u \in V$.

A3: Smoothness conditions Let $\mathbf{\Sigma}^t = [\sigma^t_{uv}]$. There exists a constant $M > 0$ such that it upper bounds the following quantities:

$$\max_{u,v \in V \times V} \sup_{t \in [0,1]} |\frac{\partial}{\partial t} \sigma^t_{uv}| < M, \quad \max_{u,v \in V \times V} \sup_{t \in [0,1]} |\frac{\partial^2}{\partial t^2} \sigma^t_{uv}| < M$$

$$\max_{u,v \in V \times V} \sup_{t \in [0,1]} |\frac{\partial}{\partial t} \theta^t_{uv}| < M, \quad \max_{u,v \in V \times V} \sup_{t \in [0,1]} |\frac{\partial^2}{\partial t^2} \theta^t_{uv}| < M.$$

The condition A3 captures our notion of the distribution that changes smoothly over time. If we consider the elements of the covariance matrix and the elements of the parameter vector as a function of time, then these functions have bounded first and second derivatives. From these assumptions, it is not too hard to see that elements of the Fisher information matrix are also smooth functions of time.

A4: Kernel The kernel $K : \mathbb{R} \mapsto \mathbb{R}$ is a symmetric function, supported in $[-1, 1]$, and there exists a constant $M_K \geq 1$ which upper bounds the quantities $\max_{z\in\mathbb{R}} |K(z)|$ and $\max_{z\in\mathbb{R}} K(z)^2$.

This condition, A4, gives some regularity conditions on the kernel used to define the weights. For example, the assumption is satisfied by the box kernel $K(z) = \frac{1}{2}\,\mathbb{1}\{z \in [-1, 1]\}$. Under the assumption A4, the kernel has the following properties:

$$2\int_{-1}^{0} zK(z)dz \leq 2\int_{-1}^{0} K(z)dz = 1$$

$$2\int_{-1}^{0} z^2K(z)dz \leq 1.$$

With the assumptions made above, we are ready to state the theorem that characterizes the consistency of the method given in the previous section for recovering the unknown time-varying graph structure. An important quantity, appearing in the statement, is the minimum value of the parameter vector that is different from zero

$$\theta_{\min} = \min_{(u,v)\in E^\tau} |\theta^\tau_{uv}|.$$

Intuitively, the success of the recovery should depend on how hard it is to distinguish the true non-zero parameters from noise.

Theorem 1. *Assume that the dependency condition A1 holds with $C_{\min}$, $D_{\min}$ and $D_{\max}$, that for each node $u \in V$, the Fisher information matrix $\mathbf{Q}^\tau$ satisfies the incoherence condition A2 with parameter α, the smoothness assumption A3 holds with parameter M, and that the kernel function used to define weights satisfies assumption A4 with parameter M_K. Let the regularization parameter satisfy*

$$\lambda_n \geq C\frac{\sqrt{\log p}}{n^{1/3}}$$

for a constant $C > 0$ independent of (n, p, s). Furthermore, assume that the following conditions hold:

1. $h = \mathcal{O}(n^{-\frac{1}{3}})$
2. $s = o(n^{1/3})$, $\frac{s^3 \log p}{n^{2/3}} = o(1)$
3. $\theta_{\min} = \Omega(\frac{\sqrt{s \log p}}{n^{1/3}})$.

Then for any $\tau \in [0,1]$, and in particular for $\tau \in \mathcal{T}_n$, the estimated graph $\hat{G}^\tau(\lambda_n)$ obtained through neighborhood selection satisfies

$$\mathbb{P}\left[\hat{G}^\tau(\lambda_n) \neq G^\tau\right] = \mathcal{O}\left(\exp\left(-C\frac{n^{2/3}}{s^3} + C'\log p\right)\right) \to 0, \tag{21}$$

for some constants C', C'' independent of (n, p, s).

This theorem guarantees that the procedure asymptotically recovers the sequence of graphs underlying all the nodal-state measurements in a time series, and the snapshot of the evolving graph at any time point during measurement intervals, under appropriate regularization parameter λ_n as long as the ambient dimensionality p and the maximum node degree s are not too large, and minimum $\boldsymbol{\theta}$ values do not tend to zero too fast. This is a somewhat surprising result because it suggests that structure recovery is possible when only one sample or even no sample exactly corresponding to the structure is available. The key insight behind this possibility is the smoothness assumption on graph evolution, which allows data points at, in theory, any time point (but in practice nearby time points determined by the kernel bandwidth) to contribute to the estimation of a graph at a particular time of interest.

Remarks:

1. The bandwidth parameter h is chosen so that it balances variance and squared bias of estimation of the elements of the Fisher information matrix.

2. Condition 2 requires that the size of the neighborhood of each node remains smaller than the size of the samples. However, the model ambient dimension p is allowed to grow exponentially in n.

3. Condition 3 is crucial to be able to distinguish true elements in the neighborhood of a node. We require that the size of the minimum element of the parameter vector stays bounded away from zero.

4. The rate of convergence is dictated by the rate of convergence of the sample Fisher information matrix to the true Fisher information matrix, as shown in Lemma 6. Using a local linear smoother, instead of the kernel smoother, to estimate the coefficients in the model (5) one could get a faster rate of convergence.

In the sequel, we set out to prove Theorem 1. The plan is to first show that the empirical estimates of the Fisher information matrix and the covariance matrix are close elementwise to their population versions. Next, we show that the minimizer $\hat{\boldsymbol{\theta}}_u^\tau$ of (17) is unique under the assumptions given in Theorem 1. Finally, we show that with high probability the estimator $\hat{\boldsymbol{\theta}}_u^\tau$ recovers the true neighborhood of a node u. Repeating the procedure for all nodes $u \in V$ we obtain the result stated in Theorem 1.

5.1.2 Large deviation inequalities

In this section we characterize the deviation of elements of the sample Fisher information matrix $\hat{\mathbf{Q}}^\tau := \hat{\mathbf{Q}}_u^\tau$ at time point τ, defined as

$$\hat{\mathbf{Q}}^\tau = \sum_t w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^\tau) \mathbf{x}_{\backslash u}^t \mathbf{x}_{\backslash u}^{t'}, \tag{22}$$

and the sample covariance matrix $\hat{\boldsymbol{\Sigma}}^\tau$ from their population versions $\mathbf{Q}^\tau$ and $\boldsymbol{\Sigma}^\tau$. As will be seen later, in the proof of the main theorem, consistency result crucially depends on the bounds on the difference $\hat{\mathbf{Q}}^\tau - \mathbf{Q}^\tau$ and $\hat{\boldsymbol{\Sigma}}^\tau - \boldsymbol{\Sigma}^\tau$. In the following, we use C, C' and C'' as generic positive constants independent of (n, p, s).

Sample Fisher information matrix

To bound the deviation between elements of $\hat{\mathbf{Q}}^\tau = [\hat{q}_{vv'}^\tau]$ and $\mathbf{Q}^\tau = [q_{vv'}^\tau]$, $v, v' \in V \backslash u$, we will use the following decomposition:

$$\begin{aligned} |\hat{q}_{vv'}^\tau - q_{vv'}^\tau| \; \leq & \; |\sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^\tau) x_v^t x_{v'}^t - \sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t| \\ & + |\sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t - \mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t]| \\ & + |\mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t] - q_{vv'}^\tau|. \end{aligned} \tag{23}$$

The following lemma gives us bounds on the terms in Eq. (23).

Lemma 2. *Assume that the smoothness condition A3 is satisfied and that the kernel function $K(\cdot)$ satisfies A4. Furthermore, assume*

$$\max_{t \in [0,1]} |\{v \in \{1, \ldots, p\} \; : \; \theta_{uv}^t \neq 0\}| < s,$$

i.e., the number of non-zero elements of the parameter vector is bounded by s. There exist constants $C, C', C'' > 0$, depending on M and M_K only, which are the constants quantifying assumption A3 and A4, respectively , such that for any $\tau \in [0, 1]$, we have

$$\max_{v,v'} \; |\hat{q}_{vv'}^\tau - \sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t| \; = \; Csh \tag{24}$$

$$\max_{v,v'} \; |\mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t] - q_{vv'}^\tau| \; = \; C'h. \tag{25}$$

Furthermore,

$$|\sum_{t \in \mathcal{T}_n} (w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t - \mathbb{E}[w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) X_v^t X_{v'}^t])| < \epsilon \tag{26}$$

with probability at least $1 - 2\exp(-C'' nh\epsilon^2)$.

Using results of Lemma 2 we can obtain the rate at which the element-wise distance between the true and sample Fisher information matrix decays to zero as a function of the bandwidth parameter h and the size of neighborhood s. In the proof of the main theorem, the bandwidth parameter will be chosen so that the bias and variance terms are balanced.

Sample covariance matrix

The deviation of the elements of the sample covariance matrix is bounded in a similar way as the deviation of elements of the sample Fisher information matrix, given in Lemma 2. Denoting the sample covariance matrix at time point τ as

$$\hat{\mathbf{\Sigma}}^{\tau} = \sum_{t} w_t^{\tau} \mathbf{x}^t \mathbf{x}^{t'}, \tag{27}$$

and the difference between the elements of $\hat{\mathbf{\Sigma}}^{\tau}$ and $\mathbf{\Sigma}^{\tau}$ can be bounded as

$$\begin{aligned} |\hat{\sigma}_{uv}^{\tau} - \sigma_{uv}^{\tau}| &= |\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t - \sigma_{uv}^{\tau}| \\ &\leq |\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t - \mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t]| \\ &+ |\mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t] - \sigma_{uv}^{\tau}|. \end{aligned} \tag{28}$$

The following lemma gives us bounds on the terms in Eq. (28).

Lemma 3. *Assume that the smoothness condition A3 is satisfied and that the kernel function $K(\cdot)$ satisfies A4. There are constants $C, C' > 0$ depending on M and M_K only such that for any $\tau \in [0,1]$, we have*

$$\max_{u,v} |\mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t] - \sigma_{uv}^{\tau}| \leq Ch. \tag{29}$$

and

$$|\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t - \mathbb{E}[\sum_{t \in \mathcal{T}_n} w_t^{\tau} x_u^t x_v^t]| \leq \epsilon \tag{30}$$

with probability at least $1 - 2\exp(-C' nh\epsilon^2)$.

A similar result was established in [34] for the case where $\mathbf{x}$ is a multivariate Normal distributed random variable.

5.1.3 Proof of Theorem 1

The proof is given through a sequence of technical lemmas. Note that in what follows, we use C, C' and C'' to denote positive constants independent of (n, p, s) and their value my change from line to line.

The main idea behind the proof is to characterize the minimum obtained in Eq. (17) and show that the correct neighborhood of one node at an arbitrary

time point can be recovered with high probability. Next, using the union bound over the nodes of a graph, we can conclude that the whole graph is estimated sparsistently at the time points of interest.

We first address the problem of uniqueness of the solution to (17). Note that because the objective in Eq. (17) is not strictly convex it is necessary to show that the non-zero pattern of the parameter vector is unique, since otherwise the problem of sparsistent graph estimation would be meaningless. Under the conditions of Theorem 1 we also have that the solution is unique, which we prove in two steps.

Let us denote the set of all solution to (17) as $\Theta(\lambda_n)$. We define the objective function in Eq. (17) by

$$F(\boldsymbol{\theta}_u) := -\sum_{t\in\mathcal{T}_n} w_t^\tau \gamma(\boldsymbol{\theta}_u; \mathbf{x}^t) + \lambda_n ||\boldsymbol{\theta}_u||_1 \tag{31}$$

and we say that $\boldsymbol{\theta}_u \in \mathbb{R}^{p-1}$ satisfies the system $(\mathcal{S})$ when

$$\forall v = 1, \ldots, p-1, \begin{cases} \sum_{t\in\mathcal{T}_n} w_t^\tau (\nabla\gamma(\boldsymbol{\theta}_u; \mathbf{x}^t))_v = \lambda_n \operatorname{sign}(\theta_{uv}) & \text{if } \theta_{uv} \neq 0 \\ |\sum_{t\in\mathcal{T}_n} w_t^\tau (\nabla\gamma(\boldsymbol{\theta}_u; \mathbf{x}^t))_v| \leq \lambda_n & \text{if } \theta_{uv} = 0, \end{cases} \tag{32}$$

where

$$\nabla\gamma(\boldsymbol{\theta}_u; \mathbf{x}^t) = \mathbf{x}^t_{\setminus u} \left\{ x_u^t + 1 - 2\mathbb{P}_{\boldsymbol{\theta}_u}[x_u^t = 1 | \mathbf{x}^t_{\setminus u}] \right\} \tag{33}$$

is the score function. Eq. (32) is obtained by taking the sub-gradient of $F(\boldsymbol{\theta})$ and equating it to zero. From the Karush-Kuhn-Tucker (KKT) conditions it follows that $\boldsymbol{\theta}_u \in \mathbb{R}^{p-1}$ belongs to $\Theta(\lambda_n)$ if and only if $\boldsymbol{\theta}_u$ satisfies the system $(\mathcal{S})$. The following Lemma shows that any two solutions have the same non-zero pattern.

Lemma 4. *Consider a node $u \in V$. If $\bar{\boldsymbol{\theta}}_u \in \mathbb{R}^{p-1}$ and $\tilde{\boldsymbol{\theta}}_u \in \mathbb{R}^{p-1}$ both belong to $\Theta(\lambda_n)$ then $\langle \mathbf{x}^t_{\setminus u}, \bar{\boldsymbol{\theta}}_u \rangle = \langle \mathbf{x}^t_{\setminus u}, \tilde{\boldsymbol{\theta}}_u \rangle$, $t \in \mathcal{T}_n$. Furthermore, solutions $\bar{\boldsymbol{\theta}}_u$ and $\tilde{\boldsymbol{\theta}}_u$ have non-zero elements in the same positions.*

We now use the result of Lemma 4 to show that with high probability the minimizer in (7) is unique. We consider the following event:

$$\Omega_{01} = \{ D_{\min} - \delta \leq \mathbf{y}' \hat{\boldsymbol{\Sigma}}^\tau_{SS} \mathbf{y} \leq D_{\max} + \delta \ : \ \mathbf{y} \in \mathbb{R}^s, ||\mathbf{y}||_2 = 1 \}.$$

Lemma 5. *Consider a node $u \in V$. Assume that the conditions of Lemma 3 are satisfied. Assume also that the dependency condition A1 holds. There are constants $C, C', C'' > 0$ depending on M and M_K only, such that*

$$\mathbb{P}[\Omega_{01}] \geq 1 - 4\exp(-Cnh(\frac{\delta}{s} - C'h)^2 + C'' \log(s)).$$

Moreover, on the event Ω_{01}, the minimizer of (7) is unique.

We have shown that the estimate $\hat{\boldsymbol{\theta}}^\tau_u$ is unique on the event Ω_{01}, which under the conditions of Theorem 1 happens with probability converging to 1 exponentially fast. To finish the proof of Theorem 1 we need to show that the

estimate $\hat{\boldsymbol{\theta}}_u^\tau$ has the same non-zero pattern as the true parameter vector $\boldsymbol{\theta}_u^\tau$. In order to show that we consider a few "good" events, which happen with high probability and on which the estimate $\hat{\boldsymbol{\theta}}_u^\tau$ has the desired properties. We start by characterizing the sample version of the Fisher information matrix, defined in Eq. (22). Consider the following events:

$$\Omega_{02} := \{C_{\min} - \delta \leq \mathbf{y}' \hat{\mathbf{Q}}_{SS}^\tau \mathbf{y} \ : \ \mathbf{y} \in \mathbb{R}^s, ||\mathbf{y}||_2 = 1\}$$

and

$$\Omega_{03} := \{\|\hat{\mathbf{Q}}_{S^cS}^\tau (\hat{\mathbf{Q}}_{SS}^\tau)^{-1}\|_\infty \leq 1 - \frac{\alpha}{2}\}.$$

Lemma 6. *Assume that the conditions of Lemma 3 are satisfied. Assume also that the dependency condition A1 holds and the incoherence condition A2 holds with the incoherence parameter α. There are constants $C, C', C'' > 0$ depending on M, M_K and α only, such that*

$$\mathbb{P}[\Omega_{02}] \geq 1 - 2\exp(-C\frac{nh\delta^2}{s^2} + C'\log(s))$$

and

$$\mathbb{P}[\Omega_{03}] \geq 1 - \exp(-C\frac{nh}{s^3} + C''\log(p)).$$

Lemma 6 guarantees that the sample Fisher information matrix satisfies "good" properties with high probability, under the appropriate scaling of quantities n, p, s and h. A similar result was obtained for the sample Fisher information matrix in [23] for the model that does not change with time. Note that the result in Lemma 6 is somewhat harder to obtain since it heavily relies on the results of Lemma 2.

We are now ready to analyze the optimum to the convex program (7). To that end we apply the mean-value theorem coordinate-wise to the gradient of the weighted logloss $\sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u; \mathbf{x}^t)$ and obtain

$$\sum_{t\in\mathcal{T}_n} w_t^\tau (\nabla\gamma(\hat{\boldsymbol{\theta}}_u^\tau; \mathbf{x}^t) - \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t)) = [\sum_{t\in\mathcal{T}_n} w_t^\tau \nabla^2\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t)](\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau) + \boldsymbol{\Delta}^\tau, \quad (34)$$

where $\boldsymbol{\Delta}^\tau \in \mathbb{R}^{p-1}$ is the remainder term of the form

$$\Delta_v^\tau \ = \ [\sum_{t\in\mathcal{T}_n} w_t^\tau (\nabla^2\gamma(\bar{\boldsymbol{\theta}}_u^{(v)}; \mathbf{x}^t) - \nabla^2\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t))]_v' (\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau) \quad (35)$$

and $\bar{\boldsymbol{\theta}}_u^{(v)}$ is a point on the line between $\boldsymbol{\theta}_u^\tau$ and $\hat{\boldsymbol{\theta}}_u^\tau$, and $[\cdot]_v'$ denoting the v-th row of the matrix. Recall that $\hat{\mathbf{Q}}^\tau = \sum_{t\in\mathcal{T}_n} w_t^\tau \nabla^2\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t)$. Using the expansion (34), we write the KKT conditions given in Eq. (32) in the following form, $\forall v = 1, \ldots, p-1$,

$$\begin{cases} \hat{\mathbf{Q}}_v^\tau(\boldsymbol{\theta}_u - \boldsymbol{\theta}_u^\tau) + \sum_{t\in\mathcal{T}_n} w_t^\tau (\nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t))_v + \Delta_v^\tau = \lambda_n \operatorname{sign}(\theta_{uv}) & \text{if } \theta_{uv} \neq 0 \\ |\hat{\mathbf{Q}}_v^\tau(\boldsymbol{\theta}_u - \boldsymbol{\theta}_u^\tau) + \sum_{t\in\mathcal{T}_n} w_t^\tau (\nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t))_v + \Delta_v^\tau| \leq \lambda_n & \text{if } \theta_{uv} = 0. \end{cases} \quad (36)$$

We consider the following events

$$\Omega_0 = \Omega_{01} \cap \Omega_{02} \cap \Omega_{03},$$

$$\Omega_1 = \{\forall v \in S \ : \ |\lambda_n((\hat{\mathbf{Q}}^\tau_{SS})^{-1}\operatorname{sign}(\boldsymbol{\theta}^\tau_S))_v - ((\hat{\mathbf{Q}}^\tau_{SS})^{-1}\mathbf{W}^\tau_S)_v| < |\theta^\tau_{uv}|\}$$

and

$$\Omega_2 = \{\forall v \in S^c \ : \ |(\mathbf{W}^\tau_{S^c} - \hat{\mathbf{Q}}^\tau_{S^cS}(\hat{\mathbf{Q}}^\tau_{SS})^{-1}\mathbf{W}^\tau_S)_v| < \frac{\alpha}{2}\lambda_n\}$$

where

$$\mathbf{W}^\tau = \sum_{t\in\mathcal{T}_n} w^\tau_t \nabla\gamma(\boldsymbol{\theta}^\tau_u; \mathbf{x}^t) + \boldsymbol{\Delta}^\tau.$$

We will work on the event Ω_0 on which the minimum eigenvalue of $\hat{\mathbf{Q}}^\tau_{SS}$ is strictly positive and, so, $\hat{\mathbf{Q}}^\tau_{SS}$ is regular and $\Omega_0 \cap \Omega_1$ and $\Omega_0 \cap \Omega_2$ are well defined.

Proposition 7. *Assume that the conditions of Lemma 6 are satisfied. The event*

$$\{\forall \hat{\boldsymbol{\theta}}^\tau_u \in \mathbb{R}^{p-1} \text{ solution of } (\mathcal{S}), \text{ we have } \operatorname{sign}(\hat{\boldsymbol{\theta}}^\tau_u) = \operatorname{sign}(\boldsymbol{\theta}^\tau_u)\} \cap \Omega_0$$

contains event $\Omega_0 \cap \Omega_1 \cap \Omega_2$.

Proof. We consider the following linear functional

$$G : \left\{ \begin{array}{ccc} \mathbb{R}^s & \to & \mathbb{R}^s \\ \boldsymbol{\theta} & \mapsto & \boldsymbol{\theta} - \boldsymbol{\theta}^\tau_S + (\hat{\mathbf{Q}}^\tau_{SS})^{-1}\mathbf{W}^\tau_S - \lambda_n(\hat{\mathbf{Q}}^\tau_{SS})^{-1}\operatorname{sign}(\boldsymbol{\theta}^\tau_S). \end{array} \right.$$

For any two vectors $\mathbf{y} = (y_1, \ldots, y_s)' \in \mathbb{R}^s$ and $\mathbf{r} = (r_1, \ldots, r_s)' \in \mathbb{R}^s_+$, define the following set centered at $\mathbf{y}$ as

$$\mathcal{B}(\mathbf{y}, \mathbf{r}) = \prod_{i=1}^{s} (y_i - r_i, y_i + r_i).$$

Now, we have

$$G\left(\mathcal{B}(\boldsymbol{\theta}^\tau_S, |\boldsymbol{\theta}^\tau_S|)\right) = \mathcal{B}\left((\hat{\mathbf{Q}}^\tau_{SS})^{-1}\mathbf{W}^\tau_S - \lambda_n(\hat{\mathbf{Q}}^\tau_{SS})^{-1}\operatorname{sign}(\boldsymbol{\theta}^\tau_S), |\boldsymbol{\theta}^\tau_S|\right).$$

On the event $\Omega_0 \cap \Omega_1$,

$$0 \in \mathcal{B}\left((\hat{\mathbf{Q}}^\tau_{SS})^{-1}\mathbf{W}^\tau_S - \lambda_n(\hat{\mathbf{Q}}^\tau_{SS})^{-1}\operatorname{sign}(\boldsymbol{\theta}^\tau_S), |\boldsymbol{\theta}^\tau_S|\right),$$

which implies that there exists a vector $\bar{\boldsymbol{\theta}}^\tau_S \in \mathcal{B}(\boldsymbol{\theta}^\tau_S, |\boldsymbol{\theta}^\tau_S|)$ such that $G(\bar{\boldsymbol{\theta}}^\tau_S) = 0$. For $\bar{\boldsymbol{\theta}}^\tau_S$ it holds that $\bar{\boldsymbol{\theta}}^\tau_S = \boldsymbol{\theta}^\tau_S + \lambda_n(\hat{\mathbf{Q}}^\tau_{SS})^{-1}\operatorname{sign}(\boldsymbol{\theta}^\tau_S) - (\hat{\mathbf{Q}}^\tau_{SS})^{-1}\mathbf{W}^\tau_S$ and $|\bar{\boldsymbol{\theta}}^\tau_S - \boldsymbol{\theta}^\tau_S| < |\boldsymbol{\theta}^\tau_S|$. Thus, the vector $\bar{\boldsymbol{\theta}}^\tau_S$ satisfies

$$\operatorname{sign}(\bar{\boldsymbol{\theta}}^\tau_S) = \operatorname{sign}(\boldsymbol{\theta}^\tau_S)$$

and

$$\hat{\mathbf{Q}}_{SS}(\bar{\boldsymbol{\theta}}^{\tau}_S - \boldsymbol{\theta}^{\tau}_S) + \mathbf{W}^{\tau}_S = \lambda_n \operatorname{sign}(\bar{\boldsymbol{\theta}}^{\tau}_S). \tag{37}$$

Next, we consider the vector $\bar{\boldsymbol{\theta}}^{\tau} = \begin{pmatrix} \bar{\boldsymbol{\theta}}^{\tau}_S \\ \bar{\boldsymbol{\theta}}^{\tau}_{S^c} \end{pmatrix}$ where $\bar{\boldsymbol{\theta}}^{\tau}_{S^c}$ is the null vector of $\mathbb{R}^{p-1-s}$. On event Ω_0, from Lemma 6 we know that $\|\hat{\mathbf{Q}}^{\tau}_{S^cS}(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1}\|_{\infty} \leq 1 - \frac{\alpha}{2}$. Now, on the event $\Omega_0 \cap \Omega_2$ it holds

$$\begin{aligned} &||\hat{\mathbf{Q}}^{\tau}_{S^cS}(\bar{\boldsymbol{\theta}}^{\tau}_S - \boldsymbol{\theta}^{\tau}_S) + \mathbf{W}^{\tau}_{S^c}||_{\infty} = \\ &\quad || - \hat{\mathbf{Q}}^{\tau}_{S^cS}(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1}\mathbf{W}^{\tau}_S + \mathbf{W}^{\tau}_{S^c} + \lambda_n \hat{\mathbf{Q}}^{\tau}_{S^cS}(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1} \operatorname{sign}(\bar{\boldsymbol{\theta}}^{\tau}_S)||_{\infty} < \lambda_n. \end{aligned} \tag{38}$$

Note that for $\bar{\boldsymbol{\theta}}^{\tau}$, equations (37) and (38) are equivalent to saying that $\bar{\boldsymbol{\theta}}^{\tau}$ satisfies conditions (36) or (32), i.e., saying that $\bar{\boldsymbol{\theta}}^{\tau}$ satisfies the KKT conditions. Since $\operatorname{sign}(\bar{\boldsymbol{\theta}}^{\tau}_S) = \operatorname{sign}(\boldsymbol{\theta}^{\tau}_S)$, we have $\operatorname{sign}(\bar{\boldsymbol{\theta}}^{\tau}) = \operatorname{sign}(\boldsymbol{\theta}^{\tau}_u)$. Furthermore, because of the uniqueness of the solution to (7) on the event Ω_0 , we conclude that $\hat{\boldsymbol{\theta}}^{\tau}_u = \bar{\boldsymbol{\theta}}^{\tau}$. □

Proposition 7 implies Theorem 1 if we manage to show that the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$ occurs with high probability under the assumptions stated in Theorem 1. Proposition 8 characterizes the probability of that event, which concludes the proof of Theorem 1.

Proposition 8. *Assume that the conditions of Theorem 1 are satisfied. Then there are constants $C, C' > 0$ depending on M, M_K, $D_{\max}$, $C_{\min}$ and α only, such that the following holds:*

$$\mathbb{P}[\Omega_0 \cap \Omega_1 \cap \Omega_2] \geq 1 - 2\exp(-Cnh(\lambda_n - sh)^2 + \log(p)). \tag{39}$$

Proof. We start the proof of the proposition by giving a technical lemma, which characterizes the distance between vectors $\hat{\boldsymbol{\theta}}^{\tau}_u = \bar{\boldsymbol{\theta}}^{\tau}$ and $\boldsymbol{\theta}^{\tau}_u$ under the assumptions of Theorem 1, where $\bar{\boldsymbol{\theta}}^{\tau}$ is constructed in the proof of Proposition 7. The following lemma gives a bound on the distance between the vectors $\hat{\boldsymbol{\theta}}^{\tau}_S$ and $\boldsymbol{\theta}^{\tau}_S$, which we use in the proof of the proposition. The proof of the lemma is given in Appendix.

Lemma 9. *Assume that the conditions of Theorem 1 are satisfied. There are constants $C, C' > 0$ depending on $M, M_K, D_{\max}, C_{\min}$ and α only, such that*

$$||\hat{\boldsymbol{\theta}}^{\tau}_S - \boldsymbol{\theta}^{\tau}_S||_2 \leq C\frac{\sqrt{s\log p}}{n^{1/3}} \tag{40}$$

with probability at least $1 - \exp(-C' \log p)$.

Using Lemma 9 we can prove Proposition 8. We start by studying the probability of the event Ω_2. We have

$$\Omega_2^C \subset \cup_{v \in S^c}\{\mathbf{W}_v + (\hat{\mathbf{Q}}^{\tau}_{S^cS}(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1}\mathbf{W}^{\tau}_S)_v \geq \frac{\alpha}{2}\lambda_n\}.$$

Recall that $\mathbf{W}^\tau = \sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t) + \boldsymbol{\Delta}^\tau$. Let us define the event

$$\Omega_3 = \{\max_{1\le v\le p-1} |\mathbf{e}_v' \sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t)| < \frac{\alpha\lambda_n}{4(2-\alpha)}\},$$

where $\mathbf{e}_v \in \mathbb{R}^{p-1}$ is a unit vector with one at the position v and zeros elsewhere. From the proof of Lemma 9 available in the appendix we have that $\mathbb{P}[\Omega_3] \ge 1 - 2\exp(-C\log(p))$ and on that event the bound given in Eq. (40) holds.

On the event Ω_3, we bound the remainder term $\boldsymbol{\Delta}^\tau$. Let $g : \mathbb{R} \mapsto \mathbb{R}$ be defined as $g(z) = \frac{4\exp(2z)}{(1+\exp(2z))^2}$. Then $\eta(\mathbf{x}; \boldsymbol{\theta}_u) = g(x_u\langle\boldsymbol{\theta}_u, \mathbf{x}_{\setminus u}\rangle)$. For $v \in \{1, \ldots, p-1\}$, using the mean value theorem it follows that

$$\begin{aligned}
\Delta_v &= [\sum_{t\in\mathcal{T}_n} w_t^\tau(\nabla^2\gamma(\bar{\boldsymbol{\theta}}_u^{(v)}; \mathbf{x}^t) - \nabla^2\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t))]_v'(\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau) \\
&= \sum_{t\in\mathcal{T}_n} w_t^\tau[\eta(\mathbf{x}^t; \bar{\boldsymbol{\theta}}_u^{(v)}) - \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^\tau)][\mathbf{x}_{\setminus u}^t \mathbf{x}_{\setminus u}^{t'}]_v'[\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau] \\
&= \sum_{t\in\mathcal{T}_n} w_t^\tau g'(\mathbf{x}_u^t\langle\bar{\bar{\boldsymbol{\theta}}}_u^{(v)}, \mathbf{x}_{\setminus u}^t\rangle)[x_u^t\mathbf{x}_{\setminus u}^t]'[\bar{\boldsymbol{\theta}}_u^{(v)} - \boldsymbol{\theta}_u^\tau][x_v^t\mathbf{x}_{\setminus u}^{t'}][\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau] \\
&= \sum_{t\in\mathcal{T}_n} w_t^\tau\{g'(x_u^t\langle\bar{\bar{\boldsymbol{\theta}}}_u^{(v)}, \mathbf{x}_{\setminus u}^t\rangle)x_u^t x_v^t\}\{[\bar{\boldsymbol{\theta}}_u^{(v)} - \boldsymbol{\theta}_u^\tau]'\mathbf{x}_{\setminus u}^t\mathbf{x}_{\setminus u}^{t'}[\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau]\},
\end{aligned}$$

where $\bar{\bar{\boldsymbol{\theta}}}_u^{(v)}$ is another point on the line joining $\hat{\boldsymbol{\theta}}_u^\tau$ and $\boldsymbol{\theta}_u^\tau$. A simple calculation shows that $|g'(x_u^t\langle\bar{\bar{\boldsymbol{\theta}}}_u^{(v)}, \mathbf{x}_{\setminus u}^t\rangle)x_u^t x_v^t| \le 1$, for all $t \in \mathcal{T}_n$, so we have

$$\begin{aligned}
|\Delta_v| &\le [\bar{\boldsymbol{\theta}}_u^{(v)} - \boldsymbol{\theta}_u^\tau]'\{\sum_{t\in\mathcal{T}_n} w_t^\tau\mathbf{x}_{\setminus u}^t\mathbf{x}_{\setminus u}^{t'}\}[\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau] \\
&\le [\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau]'\{\sum_{t\in\mathcal{T}_n} w_t^\tau\mathbf{x}_{\setminus u}^t\mathbf{x}_{\setminus u}^{t'}\}[\hat{\boldsymbol{\theta}}_u^\tau - \boldsymbol{\theta}_u^\tau] \\
&= [\hat{\boldsymbol{\theta}}_S^\tau - \boldsymbol{\theta}_S^\tau]'\{\sum_{t\in\mathcal{T}_n} w_t^\tau\mathbf{x}_S^t\mathbf{x}_S^{t'}\}[\hat{\boldsymbol{\theta}}_S^\tau - \boldsymbol{\theta}_S^\tau] \\
&\le D_{\max}||\hat{\boldsymbol{\theta}}_S^\tau - \boldsymbol{\theta}_S^\tau||_2^2.
\end{aligned} \tag{41}$$

Combining the equations (41) and (40), we have that on the event Ω_3

$$\max_{1\le v\le p-1} |\Delta_v| \le C\lambda_n^2 s < \frac{\lambda_n\alpha}{4(2-\alpha)}$$

where C is a constant depending on $D_{\max}$ and $C_{\min}$ only.

On the event $\Omega_0 \cap \Omega_3$, we have

$$W_v^\tau + (\hat{\mathbf{Q}}_{S^cS}^\tau(\hat{\mathbf{Q}}_{SS}^\tau)^{-1}\mathbf{W}_S^\tau)_v < \frac{\alpha\lambda_n}{2(2-\alpha)} + (1-\alpha)\frac{\alpha\lambda_n}{2(2-\alpha)} \le \frac{\alpha\lambda_n}{2}$$

and we can conclude that $\mathbb{P}[\Omega_2] \ge 1 - 2\exp(-C\log(p))$ for some constant C depending on $M, M_K, C_{\min}, D_{\max}$ and α only.

Next, we study the probability of the event Ω_1. We have

$$\Omega_1^C \subset \cup_{v \in S} \{\lambda_n ((\hat{\mathbf{Q}}^\tau_{SS})^{-1} \operatorname{sign}(\boldsymbol{\theta}^\tau_S))_v + ((\hat{\mathbf{Q}}^\tau_{SS})^{-1} W^\tau_S)_v \geq \theta^\tau_{uv}\}. \tag{42}$$

Again, we will consider the event Ω_3. On the event $\Omega_0 \cap \Omega_3$ we have that

$$\lambda_n ((\hat{\mathbf{Q}}^\tau_{SS})^{-1} \operatorname{sign}(\boldsymbol{\theta}^\tau_S))_v + ((\hat{\mathbf{Q}}^\tau_{SS})^{-1} \mathbf{W}^\tau_S)_v \leq \frac{\lambda_n \sqrt{s}}{C_{\min}} + \frac{\lambda_n}{2C_{\min}} \leq C\lambda_n \sqrt{s}, \tag{43}$$

for some constant C. When $\theta_{\min} > C\lambda_n \sqrt{s}$, we have that $\mathbb{P}[\Omega_1] \geq 1 - 2\exp(-C\log(p))$ for some constant C that depends on $M, M_K, C_{\min}, D_{\max}$ and α only. □

In summary, under the assumptions of Theorem 1, the probability of event $\Omega_0 \cap \Omega_1 \cap \Omega_2$ converges to one exponentially fast. On this event, we have shown that the estimator $\hat{\boldsymbol{\theta}}^\tau_u$ is the unique minimizer of (17) and that it consistently estimates the signed non-zero pattern of the true parameter vector $\boldsymbol{\theta}^\tau_u$, i.e., it consistently estimates the neighborhood of a node u. Applying the union bound over all nodes $u \in V$, we can conclude that our estimation procedure consistently estimates the graph structure at a time point τ.

5.2 Recovery under structural changes

Currently we do not have a consistency result for the estimator produced by the method `TV`, however, we have obtained some insight on how to solve this problem and plan to pursue it in our future research. The main difficulty seems to be the presence of both the ℓ_1 and TV$(\cdot)$ regularization terms in Eq. (10), which complicates the analysis. However, if we relate the method `TV` to the problem of multiple change point detection, we can observe the following: the TV$(\cdot)$ penalty biases the estimate $\{\hat{\boldsymbol{\theta}}^t\}_{t \in \mathcal{T}_n}$ towards a piecewise constant solution, and this effectively partitions the time interval $[0, 1]$ into segments within which the parameter is constant. If we can estimate the partition $\mathcal{B}_u$ correctly, then the graph structure can also be estimated successfully if there are enough samples on each segment of the partition. In fact, [25] observed that it is useful to consider a two-stage procedure in which the first stage uses the total variation penalty to estimate the partition, and the second stage then uses the ℓ_1 penalty to determine non-zero parameters within each segment. Although his analysis is restricted to the fused lasso [28], we believe that his techniques can be extended for analyzing our method `TV`. Besides assumptions 1 to 4 which appeared in method `smooth`, additional assumptions may be needed to assure the consistent estimation of the partition $\mathcal{B}_u$.

Partial results along this direction were presented in [18] for linear regression model. Some additional work is needed to adapt the proof technique to the case of logistic regression.

6 Discussion

We have presented two algorithms for an important problem of structure estimation of time varying networks. While the structure estimation of the static networks is an important problem in itself, in certain cases static structures are of limited use. More specifically, a static structure only shows connections and interactions that are persistent throughout the whole time period, and therefore time varying structures are needed to describe dynamic interactions that are transient in time. Although the algorithms presented in this paper for learning time varying networks are simple, they can already be used to discover some patterns that would not be discovered using a method that estimates static networks. However, the ability to learn time varying networks comes at a price of extra tuning parameters: the bandwidth parameter h or the penalty parameter λ_{TV}.

Throughout the paper, we assume that the observations at different points in time are independent. An important future direction is the analysis of the graph structure estimation from a general time-series, with dependent observations. In our opinion, this extension will be straightforward but with great practical importance. Furthermore, we have worked with the assumption that the data are binary, however, extending the procedure to work with multi-category data is also straightforward. One possible approach is explained in [23] and can be directly used here.

There are still ways to improve the methods presented here. For instance, more principled ways of selecting tuning parameters are definitely needed. Selecting the tuning parameters in neighborhood selection procedure for static graphs is not an easy problem, and estimating time varying graphs makes the problem more challenging. Furthermore, methods presented here do not allow for the incorporation of existing knowledge on the network topology into the algorithm. In some cases, the data are very scarce and we would like to incorporate as much prior knowledge as possible, so developing Bayesian methods seems very important.

The method `smooth` and the method `TV` represent two different ends of the spectrum: one algorithm is able to estimate smoothly changing networks, while the other one is tailored towards estimation of structural changes in the model. It is important to bring the two methods together in the future work. There is a great amount of work on nonparametric estimation of change points and it would be interesting to incorporate those methods for estimating time varying networks.

7 Appendix

Note that in what follows, we use C, C' and C'' to denote positive constants and their value may change from line to line.

7.1 Proof of Lemma 2

We start the proof by bounding the difference $|\eta(\mathbf{x};\boldsymbol{\theta}_u^{t+\delta}) - \eta(\mathbf{x};\boldsymbol{\theta}_u^t)|$ which will be useful later on. By applying the mean value theorem to $\eta(\mathbf{x};\cdot)$ and the Taylor expansion on $\boldsymbol{\theta}_u^t$ we obtain:

$$
\begin{aligned}
|\eta(\mathbf{x};\boldsymbol{\theta}_u^{t+\delta}) - \eta(\mathbf{x};\boldsymbol{\theta}_u^t)| &= |\sum_{v=1}^{p-1}(\theta_{uv}^{t+\delta} - \theta_{uv}^t)\eta'(\mathbf{x};\bar{\boldsymbol{\theta}}_u^{(v)})| \qquad \left(\begin{array}{c}\bar{\boldsymbol{\theta}}_u^{(v)} \text{ is a point on the line} \\ \text{between } \boldsymbol{\theta}_u^{t+\delta} \text{ and } \boldsymbol{\theta}_u^t\end{array}\right) \\
&\le \sum_{v=1}^{p-1}|\theta_{uv}^{t+\delta} - \theta_{uv}^t| \qquad (\ |\eta'(\mathbf{x};\cdot)| \le 1\ ) \\
&= \sum_{v=1}^{p-1}|\delta\frac{\partial}{\partial t}\theta_{uv}^t + \frac{\delta^2}{2}\frac{\partial^2}{\partial t^2}\theta_{uv}^t\Big|_{t=\beta_v}| \qquad \left(\begin{array}{c}\beta_v \text{ is a point on the line} \\ \text{between } t \text{ and } t+\delta\end{array}\right)
\end{aligned}
$$

Without loss of generality, let $\tau = 1$. Using the above equation, and the Riemann integral to approximate the sum, we have

$$
\begin{aligned}
&|\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^\tau)x_v^t x_{v'}^t - \sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^t)x_v^t x_{v'}^t| \\
&\approx |\int \frac{2}{h}K(\frac{z-\tau}{h})[\eta(\mathbf{x}^z;\boldsymbol{\theta}_u^\tau) - \eta(\mathbf{x}^z;\boldsymbol{\theta}_u^z)]x_v^z x_{v'}^z dz| \\
&\le 2\int_{-\frac{1}{h}}^0 K(z')|\eta(\mathbf{x}^{\tau+z'h};\boldsymbol{\theta}_u^\tau) - \eta(\mathbf{x}^{\tau+z'h};\boldsymbol{\theta}_u^{\tau+z'h})|dz' \\
&\le 2\int_{-1}^0 K(z')[\sum_{v=1}^{p-1}|z'h\frac{\partial}{\partial t}\theta_{uv}^t\Big|_{t=\tau} + \frac{(z'h)^2}{2}\frac{\partial^2}{\partial t^2}\theta_{uv}^t\Big|_{t=\beta_v}|]dz' \\
&\le Csh,
\end{aligned}
$$

for some constant $C > 0$ depending on M from A3 which bounds the derivatives in the equation above, and M_K from A4 which bounds the kernel. The last inequality follows from the assumption that the number of non-zero components of the vector $\boldsymbol{\theta}_u^t$ is bounded by s.

Next, we prove equation (25). Using the Taylor expansion, for any fixed $1 \le v, v' \le p-1$ we have

$$
\begin{aligned}
&|\mathbb{E}[\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^t)x_v^t x_{v'}^t] - q_{vv'}^\tau| \\
&= |\sum_{t\in\mathcal{T}_n} w_t^\tau (q_{vv'}^t - q_{vv'}^\tau)| \\
&= |\sum_{t\in\mathcal{T}_n} w_t^\tau ((t-\tau)\frac{\partial}{\partial t}q_{vv'}^t\Big|_{t=\tau} + \frac{(t-\tau)^2}{2}\frac{\partial^2}{\partial t^2}q_{vv'}^t\Big|_{t=\xi}|,
\end{aligned}
$$

where $\xi \in [t,\tau]$. Since $w_t^\tau = 0$ for $|t-\tau| > h$, we have

$$
\max_{v,v'} |\mathbb{E}[\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^t)x_v^t x_{v'}^t] - q_{vv'}^\tau| \le C'h
$$

for some constant $C > 0$ depending on M and M_K only.

Finally, we prove equation (26). Observe that $w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^t) x_v^t x_{v'}^t$ are independent and bounded random variables $[-w_t^\tau, w_t^\tau]$. The equation simply follows from the Hoeffding's inequality. □

7.2 Proof of Lemma 3

To obtain the Lemma, we follow the same proof strategy as in the proof of Lemma 2. In particular, Eq. (29) is proved in the same way as Eq. (25) and Eq. (30) in the same way as Eq. (26). The details of this derivation are omitted. □

7.3 Proof of Lemma 4

The set of minima $\Theta(\lambda_n)$ of a convex function is convex. So, for two distinct points of minima, $\bar{\boldsymbol{\theta}}_u$ and $\tilde{\boldsymbol{\theta}}_u$, every point on the line connecting two points also belongs to minima, i.e. $\xi\bar{\boldsymbol{\theta}}_u + (1-\xi)\tilde{\boldsymbol{\theta}}_u \in \Theta(\lambda_n)$, for any $\xi \in (0,1)$. Let $\boldsymbol{\eta} = \bar{\boldsymbol{\theta}}_u - \tilde{\boldsymbol{\theta}}_u$ and now any point on the line can be written as $\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}$. The value of the objective at any point of minima is constant and we have

$$F(\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}) = c, \quad \xi \in (0,1),$$

where c is some constant. By taking the derivative with respect to ξ of $F(\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta})$ we obtain

$$\begin{aligned}\sum_{t\in\mathcal{T}_n} w_t^\tau \left[-x_u^t + \frac{\exp(\langle\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle) - \exp(-\langle\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle)}{\exp(\langle\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle) + \exp(-\langle\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle)}\right] \langle\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle \\ + \lambda_n \sum_{v=1}^{p-1} \eta_v \operatorname{sign}(\tilde{\theta}_{uv} + \xi\eta_v) = 0.\end{aligned} \tag{44}$$

On a small neighborhood of ξ the sign of $\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}$ is constant, for each component v, since the function $\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}$ is continuous in ξ. By taking the derivative with respect to ξ of Eq. (44) and noting that the last term is constant on a small neighborhood of ξ we have

$$4\sum_{t\in\mathcal{T}_n} w_t^\tau \langle\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle^2 \frac{\exp(-2\langle\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle)}{\left(1 + \exp(-2\langle\tilde{\boldsymbol{\theta}}_u + \xi\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle)\right)^2} = 0.$$

This implies that $\langle\boldsymbol{\eta}, \mathbf{x}_{\setminus u}^t\rangle = 0$ for every $t \in \mathcal{T}_n$, which implies that $\langle\mathbf{x}_{\setminus u}^t, \bar{\boldsymbol{\theta}}_u\rangle = \langle\mathbf{x}_{\setminus u}^t, \tilde{\boldsymbol{\theta}}_u\rangle$, $t \in \mathcal{T}_n$, for any two solutions $\bar{\boldsymbol{\theta}}_u$ and $\tilde{\boldsymbol{\theta}}_u$. Since $\bar{\boldsymbol{\theta}}_u$ and $\tilde{\boldsymbol{\theta}}_u$ were two arbitrary elements of $\Theta(\lambda_n)$ we can conclude that $\langle\mathbf{x}_{\setminus u}^t, \boldsymbol{\theta}_u\rangle$, $t \in \mathcal{T}_n$ is constant for all elements $\boldsymbol{\theta}_u \in \Theta(\lambda_n)$.

Next, we need to show that the conclusion from above implies that any two solutions have non-zero elements in the same position. From equation (32), it

follows that the set of non-zero components of the solution is given by

$$S = \left\{ 1 \leq v \leq p-1 \ : \ \left| \sum_{t \in \mathcal{T}_n} w_t^\tau (\nabla \gamma(\boldsymbol{\theta}_u; \mathbf{x}^t))_v \right| = \lambda \right\}.$$

Using equation (33) we have that

$$\sum_{t \in \mathcal{T}_n} w_t^\tau (\nabla \gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t))_v =$$
$$\sum_{t \in \mathcal{T}_n} w_t^\tau (\mathbf{x}^t_{\setminus u} \{ x_u^t + 1 - 2 \frac{\exp(2 x_u^t \langle \boldsymbol{\theta}_u^\tau, \mathbf{x}^t_{\setminus u} \rangle)}{\exp(2 x_u^t \langle \boldsymbol{\theta}_u^\tau, \mathbf{x}^\tau_{\setminus u} \rangle) + 1} \})_v,$$

which is constant across different elements $\boldsymbol{\theta}_u \in \Theta(\lambda_n)$, since $\langle \mathbf{x}^t_{\setminus u}, \boldsymbol{\theta}_u \rangle$, $t \in \mathcal{T}_n$ is constant for all $\boldsymbol{\theta}_u \in \Theta(\lambda_n)$. This implies that the set of non-zero components is the same for all solutions. □

7.4 Proof of Lemma 5

Under the assumptions given in the Lemma, we can apply the result of Lemma 3. Let $\mathbf{y} \in \mathbb{R}^s$ be a unit norm minimal eigenvector of $\hat{\boldsymbol{\Sigma}}^\tau_{SS}$. We have

$$\begin{aligned} \Lambda_{\min}(\boldsymbol{\Sigma}^\tau_{SS}) &= \min_{||\mathbf{x}||_2 = 1} \mathbf{x}' \boldsymbol{\Sigma}^\tau_{SS} \mathbf{x} \\ &= \min_{||\mathbf{x}||_2 = 1} \{ \mathbf{x}' \hat{\boldsymbol{\Sigma}}^\tau_{SS} \mathbf{x} + \mathbf{x}' (\boldsymbol{\Sigma}^\tau_{SS} - \hat{\boldsymbol{\Sigma}}^\tau_{SS}) \mathbf{x} \} \\ &\leq \mathbf{y}' \hat{\boldsymbol{\Sigma}}^\tau_{SS} \mathbf{y} + \mathbf{y}' (\boldsymbol{\Sigma}^\tau_{SS} - \hat{\boldsymbol{\Sigma}}^\tau_{SS}) \mathbf{y}, \end{aligned}$$

which implies

$$\Lambda_{\min}(\hat{\boldsymbol{\Sigma}}^\tau_{SS}) \geq D_{\min} - |||(\boldsymbol{\Sigma}^\tau_{SS} - \hat{\boldsymbol{\Sigma}}^\tau_{SS})|||_2.$$

Let $\boldsymbol{\Sigma}^\tau = [\sigma^\tau_{uv}]$ and $\hat{\boldsymbol{\Sigma}}^\tau = [\hat{\sigma}^\tau_{uv}]$. We have the following bound on the spectral norm

$$|||\boldsymbol{\Sigma}^\tau_{SS} - \hat{\boldsymbol{\Sigma}}^\tau_{SS}|||_2 \leq \left(\sum_{u=1}^s \sum_{v=1}^s (\hat{\sigma}^\tau_{uv} - \sigma^\tau_{uv})^2 \right)^{1/2} \leq \delta,$$

with the probability at least $1 - 2\exp(-Cnh(\frac{\delta}{s} - C'h)^2 + C'' \log(s))$, for some fixed constants $C, C', C'' > 0$ depending on M and M_K only.

Similarly, we have that

$$\Lambda_{\max}(\hat{\boldsymbol{\Sigma}}^\tau_{SS}) \leq D_{\max} + \delta,$$

with probability at least $1 - 2\exp(-Cnh(\frac{\delta}{s} - C'h)^2 + C'' \log(s))$, for some fixed constants $C, C', C'' > 0$ depending on M and M_K only.

From Lemma 4, we know that any two solutions $\bar{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\theta}}_u \in \Theta(\lambda_n)$ of the optimization problem (17) have non-zero elements in the same position. So, for any two solutions $\bar{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\theta}}_u \in \Theta(\lambda_n)$, it holds

$$\mathbf{X}_{\setminus u}(\bar{\boldsymbol{\theta}}_u - \tilde{\boldsymbol{\theta}}_u) = \mathbf{X}_{\setminus u, S}(\bar{\boldsymbol{\theta}}_u - \tilde{\boldsymbol{\theta}}_u)_S + \mathbf{X}_{\setminus u, S^c}(\bar{\boldsymbol{\theta}}_u - \tilde{\boldsymbol{\theta}}_u)_{S^c} = \mathbf{X}_{\setminus u, S}(\bar{\boldsymbol{\theta}}_u - \tilde{\boldsymbol{\theta}}_u)_S.$$

Furthermore, from Lemma 4 we know that the two solutions are in the kernel of $\mathbf{X}_{\setminus u,S}$. On the event Ω_{01}, kernel of $\mathbf{X}_{\setminus u,S}$ is $\{0\}$. Thus, the solution is unique on Ω_{01}. □

7.5 Proof of Lemma 6

We first analyze the probability of the event Ω_{02}. Using the same argument to those in the proof of Lemma 5, we obtain

$$\Lambda_{\min}(\hat{\mathbf{Q}}^{\tau}_{SS}) \geq C_{\min} - |||\mathbf{Q}^{\tau}_{SS} - \hat{\mathbf{Q}}^{\tau}_{SS}|||_2.$$

Next, using results of Lemma 2, we have the following bound

$$|||\mathbf{Q}^{\tau}_{SS} - \hat{\mathbf{Q}}^{\tau}_{SS}|||_2 \leq \left(\sum_{u=1}^{s}\sum_{v=1}^{s}(\hat{q}^{\tau}_{uv} - q^{\tau}_{uv})^2\right)^{1/2} \leq \delta, \tag{45}$$

with probability at least $1 - 2\exp(-C\frac{nh\delta^2}{s^2} + 2\log(s))$, for some fixed constants $C, C' > 0$ depending on M and M_K only.

Next, we deal with the event Ω_{03}. We are going to use the following decomposition

$$\begin{aligned}
\hat{\mathbf{Q}}^{\tau}_{S^cS}(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1} &= \mathbf{Q}^{\tau}_{S^cS}[(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1} - (\mathbf{Q}^{\tau}_{SS})^{-1}] \\
&+ [\hat{\mathbf{Q}}^{\tau}_{S^cS} - \mathbf{Q}^{\tau}_{S^cS}](\mathbf{Q}^{\tau}_{SS})^{-1} \\
&+ [\hat{\mathbf{Q}}^{\tau}_{S^cS} - \mathbf{Q}^{\tau}_{S^cS}][(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1} - (\mathbf{Q}^{\tau}_{SS})^{-1}] \\
&+ \mathbf{Q}^{\tau}_{S^cS}(\mathbf{Q}^{\tau}_{SS})^{-1} \\
&= T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

Under the assumption A2, we have that $|||T_4|||_\infty \leq 1 - \alpha$. The lemma follows if we prove that for all the other terms we have $|||\cdot|||_\infty \leq \frac{\alpha}{6}$. Using the submultiplicative property of the norm, we have for the first term:

$$\begin{aligned}
|||T_1|||_\infty &\leq |||\mathbf{Q}^{\tau}_{S^cS}\left(\mathbf{Q}^{\tau}_{SS}\right)^{-1}|||_\infty |||\hat{\mathbf{Q}}^{\tau}_{SS} - \mathbf{Q}^{\tau}_{SS}|||_\infty |||(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1}|||_\infty \\
&\leq (1-\alpha)|||\hat{\mathbf{Q}}^{\tau}_{SS} - \mathbf{Q}^{\tau}_{SS}|||_\infty \sqrt{s} |||(\hat{\mathbf{Q}}^{\tau}_{SS})^{-1}|||_2.
\end{aligned} \tag{46}$$

Using Eq. (45), we can bound the term $|||\left(\hat{\mathbf{Q}}^{\tau}_{SS}\right)^{-1}|||_2 \leq C''$, for some constant depending on $C_{\min}$ only, with probability at least $1 - 2\exp(-C\frac{nh}{s} + 2\log(s))$, for some fixed constant $C > 0$. The bound on the term $|||\hat{\mathbf{Q}}^{\tau}_{SS} - \mathbf{Q}^{\tau}_{SS}|||_\infty$ follows from application of Lemma 2. Observe that

$$\begin{aligned}
\mathbb{P}[|||\hat{\mathbf{Q}}^{\tau}_{SS} - \mathbf{Q}^{\tau}_{SS}|||_\infty \geq \delta] &= \mathbb{P}[\max_{v\in S}\{\sum_{v'\in S}|\hat{q}^{\tau}_{vv'} - q^{\tau}_{vv'}|\} \geq \delta] \\
&\leq 2\exp(-Cnh(\frac{\delta}{s} - C'sh)^2 + 2\log(s)),
\end{aligned} \tag{47}$$

for some fixed constants $C, C' > 0$. Combining all the elements, we obtain the bound on the first term $\|T_1\|_\infty \leq \frac{\alpha}{6}$, with probability at least $1 - C\exp(C'\frac{nh}{s^3} + C''\log(s))$, for some constants $C, C', C'' > 0$.

Next, we analyze the second term. We have that

$$\begin{aligned} \|T_2\|_\infty &\leq \|\hat{\mathbf{Q}}^\tau_{S^cS} - \mathbf{Q}^\tau_{S^cS}\|_\infty \sqrt{s} \| \left(\mathbf{Q}^\tau_{SS}\right)^{-1} \|_2 \\ &\leq \frac{\sqrt{s}}{C_{\min}} \|\hat{\mathbf{Q}}^\tau_{S^cS} - \mathbf{Q}^\tau_{S^cS}\|_\infty. \end{aligned} \tag{48}$$

The bound on the term $\|\hat{\mathbf{Q}}^\tau_{SS} - \mathbf{Q}^\tau_{SS}\|_\infty$ follows in the same way as the bound in Eq. (47) and we can conclude that $\|T_3\|_\infty \leq \frac{\alpha}{6}$ with probability at least $1 - C\exp(C'\frac{nh}{s^3} + C''\log(p))$, for some constants $C, C', C'' > 0$.

Finally, we bound the third term T_3. We have the following decomposition

$$\begin{aligned} &\|[\hat{\mathbf{Q}}^\tau_{S^cS} - \mathbf{Q}^\tau_{S^cS}][(\hat{\mathbf{Q}}^\tau_{SS})^{-1} - (\mathbf{Q}^\tau_{SS})^{-1}]\|_\infty \\ &\quad\leq \|\hat{\mathbf{Q}}^\tau_{S^cS} - \mathbf{Q}^\tau_{S^cS}\|_\infty \sqrt{s} \|(\mathbf{Q}^\tau_{SS})^{-1}[\mathbf{Q}^\tau_{SS} - \hat{\mathbf{Q}}^\tau_{SS}](\hat{\mathbf{Q}}^\tau_{SS})^{-1}\|_2 \\ &\quad\leq \frac{\sqrt{s}}{C_{\min}} \|\hat{\mathbf{Q}}^\tau_{S^cS} - \mathbf{Q}^\tau_{S^cS}\|_\infty \|\mathbf{Q}^\tau_{SS} - \hat{\mathbf{Q}}^\tau_{SS}\|_2 \|(\hat{\mathbf{Q}}^\tau_{SS})^{-1}\|_2. \end{aligned}$$

Bounding the remaining terms as in equations (48), (47) and (46), we obtain that $\|T_3\|_\infty \leq \frac{\alpha}{6}$ with probability at least $1 - C\exp(C'\frac{nh}{s^3} + C''\log(p))$.

Bound on the probability of event Ω_{03} follows from combining the bounds on all terms. □

7.6 Proof of Lemma 9

To prove this Lemma, we use a technique of Rothman et al. [26] applied to the problem of consistency of the penalized covariance matrix estimator. Let us define the following function

$$H : \left\{ \begin{array}{ccc} \mathbb{R}^p & \rightarrow & \mathbb{R} \\ \mathbf{D} & \mapsto & F(\boldsymbol{\theta}^\tau_u + \mathbf{D}) - F(\boldsymbol{\theta}^\tau_u), \end{array} \right.$$

where the function $F(\cdot)$ is defined in equation (31). The function $H(\cdot)$ takes the following form

$$\begin{aligned} H(\mathbf{D}) = \sum_{t \in \mathcal{T}_n} & w^\tau_t (\gamma(\boldsymbol{\theta}^\tau_u; \mathbf{x}^t) - \gamma(\boldsymbol{\theta}^\tau_u + \mathbf{D}; \mathbf{x}^t)) \\ &+ \lambda_n(||\boldsymbol{\theta}^\tau_u + \mathbf{D}||_1 - ||\boldsymbol{\theta}^\tau_u||_1). \end{aligned}$$

Recall the minimizer of (7) constructed in the proof of Proposition 7, $\hat{\boldsymbol{\theta}}^\tau_u = (\bar{\boldsymbol{\theta}}'_S, 0'_{S^c})'$. The minimizer of the function $H(\cdot)$ is $\hat{\mathbf{D}} = \hat{\boldsymbol{\theta}}^\tau_u - \boldsymbol{\theta}^\tau_u$. Function $H(\cdot)$ is convex and $H(0) = 0$ by construction. Therefor $H(\hat{\mathbf{D}}) \leq 0$. If we show that for some radius $B > 0$, and $\mathbf{D} \in \mathbb{R}^p$ with $||\mathbf{D}||_2 = B$ and $\mathbf{D}_{S^c} = \mathbf{0}$, we have $H(\mathbf{D}) > 0$, then we claim that $||\hat{\mathbf{D}}||_2 \leq B$. This follows from the convexity of $H(\cdot)$.

We proceed to show strict positivity of $H(\cdot)$ on the boundary of the ball with radius $B = K\lambda_n\sqrt{s}$, where $K > 0$ is a parameter to be chosen wisely later. Let $\mathbf{D} \in \mathbb{R}^p$ be an arbitrary vector with $||\mathbf{D}||_2 = B$ and $\mathbf{D}_{S^c} = \mathbf{0}$, then by the Taylor expansion of $\gamma(\cdot; \mathbf{x}^t)$ we have

$$\begin{aligned} H(\mathbf{D}) &= -(\sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t))'\mathbf{D} \\ &\quad - \mathbf{D}'[\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t; \boldsymbol{\theta}_u^\tau + \alpha\mathbf{D})\mathbf{x}_{\setminus u}^t \mathbf{x}_{\setminus u}^{t'}]\mathbf{D} \\ &\quad + \lambda_n(||\boldsymbol{\theta}_u^\tau + \mathbf{D}||_1 - ||\boldsymbol{\theta}_u^\tau||_1) \\ &= (I) + (II) + (III), \end{aligned} \tag{49}$$

for some $\alpha \in [0, 1]$.

We start from the term (I). Let $\mathbf{e}_v \in \mathbb{R}^p$ be a unit vector with one at the position v and zeros elsewhere. Then random variables $-\mathbf{e}_v' \sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t)$ are bounded $[-\frac{C}{nh}, \frac{C}{nh}]$ for all $1 \le v \le p-1$, with constant $C > 0$ depending on M_K only. Using the Hoeffding inequality and the union bound, we have

$$\max_{1\le v\le p-1} |\mathbf{e}_v'(\sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t) - \mathbb{E}[\sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t)])| \le \delta,$$

with probability at least $1 - 2\exp(-Cnh\delta^2 + \log(p))$, where $C > 0$ is a constant depending on M_K only. Moreover, denoting

$$p(\boldsymbol{\theta}_u^t) = \mathbb{P}_{\boldsymbol{\theta}_u^t}[x_u^t = 1 \mid \mathbf{x}_{\setminus u}^t]$$

to simplify the notation, we have for all $1 \le v \le p-1$,

$$\begin{aligned} &|\mathbb{E}[\mathbf{e}_v' \sum_{t\in\mathcal{T}_n} w_t^\tau \nabla\gamma(\boldsymbol{\theta}_u^\tau; \mathbf{x}^t) \mid \{\mathbf{x}_{\setminus u}^t\}_{t\in\mathcal{T}_n}]| \\ &= |\mathbb{E}[\sum_{t\in\mathcal{T}_n} w_t^\tau x_v^t [x_u^t + 1 - 2p(\boldsymbol{\theta}_u^\tau)] \mid \{\mathbf{x}_{\setminus u}^t\}_{t\in\mathcal{T}_n}]| \\ &= |2\sum_{t\in\mathcal{T}_n} w_t^\tau x_v^t [p(\boldsymbol{\theta}_u^t) - p(\boldsymbol{\theta}_u^\tau)]| \\ &\le 4\int_{-\frac{1}{h}}^{0} K(z)|p(\boldsymbol{\theta}_u^{\tau+zh}) - p(\boldsymbol{\theta}_u^\tau)|dz. \end{aligned} \tag{50}$$

Next, we apply the mean value theorem on $p(\cdot)$ and the Taylor's theorem on

$\boldsymbol{\theta}_u^t$. Under the assumption A3, we have

$$\begin{aligned}
&|p(\boldsymbol{\theta}_u^{\tau+zh}) - p(\boldsymbol{\theta}_u^{\tau})| \\
&\quad\leq \sum_{v=1}^{p-1} |\boldsymbol{\theta}_{uv}^{\tau+zh} - \theta_{uv}^{\tau}| \qquad (|\, p'(\cdot)| \leq 1\ ) \\
&\quad= \sum_{v=1}^{p-1} |zh \frac{\partial}{\partial t}\theta_{uv}^t\Big|_{t=\tau} + \frac{(zh)^2}{2}\frac{\partial^2}{\partial t^2}\theta_{uv}^t\Big|_{t=\alpha_v}| \qquad (\ \alpha_v \in [\tau+zh, \tau]\ ) \\
&\quad\leq Cs|zh + \frac{(zh)^2}{2}|,
\end{aligned} \tag{51}$$

for some $C > 0$ depending only on M. Combining (51) and (50) we have that $|\mathbb{E}[\mathbf{e}_v' \sum_{t\in\mathcal{T}_n} w_t^{\tau}\nabla\gamma(\boldsymbol{\theta}_u^{\tau}; \mathbf{x}^t)| \leq Csh$ for all $1 \leq v \leq p-1$. Thus, with probability greater than $1 - 2\exp(-Cnh(\lambda_n - sh)^2 + \log(p))$ for some constant $C > 0$ depending only on M_K, M and α, which under the conditions of Theorem 1 goes to 1 exponentially fast, we have

$$\max_{1\leq v\leq p-1} |\mathbf{e}_v' \sum_{t\in\mathcal{T}_n} w_t^{\tau}\nabla\gamma(\boldsymbol{\theta}_u^{\tau}; \mathbf{x}^t)| \leq \frac{\alpha\lambda_n}{4(2-\alpha)} < \frac{\lambda_n}{4}.$$

On that event, using Hölder's inequality, we have

$$\begin{aligned}
|(\sum_{t\in\mathcal{T}_n} w_t^{\tau}\nabla\gamma(\boldsymbol{\theta}_u^{\tau}; \mathbf{x}^t))'\mathbf{D}| &\leq ||\mathbf{D}||_1 \max_{1\leq v\leq p-1} |\mathbf{e}_v' \sum_{t\in\mathcal{T}_n} w_t^{\tau}\nabla\gamma(\boldsymbol{\theta}_u^{\tau}; \mathbf{x}^t)| \\
&\leq \frac{\lambda_n}{4}\sqrt{s}||\mathbf{D}||_2 \leq (\lambda_n\sqrt{s})^2\frac{K}{4}.
\end{aligned}$$

The triangle inequality applied to the term (III) of equation (49) yields:

$$\begin{aligned}
\lambda_n(||\boldsymbol{\theta}_u^{\tau} + \mathbf{D}||_1 - ||\boldsymbol{\theta}_u^{\tau}||_1) &\geq -\lambda_n||\mathbf{D}_S||_1 \\
&\geq -\lambda_n\sqrt{s}||\mathbf{D}_S||_2 \geq -K(\lambda_n\sqrt{s})^2.
\end{aligned}$$

Finally, we bound the term (II) of equation (49). Observe that since $\mathbf{D}_{S^c} = 0$, we have

$$\begin{aligned}
&\mathbf{D}'[\sum_{t\in\mathcal{T}_n} w_t^{\tau}\eta(\mathbf{x}^t; \boldsymbol{\theta}_u^{\tau} + \alpha\mathbf{D})\mathbf{x}_{\setminus u}^t\mathbf{x}_{\setminus u}^{t'}]\mathbf{D} \\
&\quad= \mathbf{D}_S'[\sum_{t\in\mathcal{T}_n} w_t^{\tau}\eta(\mathbf{x}^t; \boldsymbol{\theta}_u^{\tau} + \alpha\mathbf{D})\mathbf{x}_S^t\mathbf{x}_S^{t'}]\mathbf{D}_S \\
&\quad\geq K^2\Lambda_{\min}(\sum_{t\in\mathcal{T}_n} w_t^{\tau}\eta(\mathbf{x}^t; \boldsymbol{\theta}_u^{\tau} + \alpha\mathbf{D})\mathbf{x}_S^t\mathbf{x}_S^{t'})
\end{aligned}$$

Let $g : \mathbb{R} \mapsto \mathbb{R}$ be defined as $g(z) = \frac{4\exp(2z)}{(1+\exp(2z))^2}$. Now, $\eta(\mathbf{x}; \boldsymbol{\theta}_u) = g(x_u\langle\boldsymbol{\theta}_u, \mathbf{x}_{\setminus u}\rangle)$

and we have

$$\begin{aligned}
&\Lambda_{\min}(\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D})\mathbf{x}_S^t\mathbf{x}_S^{t'}) \\
&\quad\geq \min_{\alpha\in[0,1]}\Lambda_{\min}(\sum_{t\in\mathcal{T}_n} w_t\eta(\mathbf{x}^t;\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D})\mathbf{x}_S^t\mathbf{x}_S^{t'}) \\
&\quad\geq \Lambda_{\min}(\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^\tau)\mathbf{x}_S^t\mathbf{x}_S^{t'}) \\
&\qquad - \max_{\alpha\in[0,1]}\|\sum_{t\in\mathcal{T}_n} w_t^\tau g'(x_u^t\langle\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D},\mathbf{x}_S^t\rangle)(x_u^t\mathbf{D}'_S\mathbf{x}_S^t)\mathbf{x}_S^t\mathbf{x}_S^{t'}\|_2 \\
&\quad\geq C_{\min} - \max_{\alpha\in[0,1]}\|\sum_{t\in\mathcal{T}_n} w_t^\tau g'(x_u^t\langle\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D},\mathbf{x}_S^t\rangle)(x_u^t\mathbf{D}'_S\mathbf{x}_S^t)\mathbf{x}_S^t\mathbf{x}_S^{t'}\|_2
\end{aligned}$$

To bound the spectral norm, we observe that for any fixed $\alpha\in[0,1]$ and $y\in\mathbb{R}^s, ||\mathbf{y}||_2=1$ we have:

$$\begin{aligned}
&\mathbf{y}'\{\sum_{t\in\mathcal{T}_n} w_t^\tau g'(x_u^t\langle\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D},\mathbf{x}_S^t\rangle)(x_u^t\mathbf{D}'_S\mathbf{x}_S^t)\mathbf{x}_S^t\mathbf{x}_S^{t'}\}\mathbf{y} \\
&\quad= \sum_{t\in\mathcal{T}_n} w_t^\tau g'(x_u^t\langle\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D},\mathbf{x}_S^t\rangle)(x_u^t\mathbf{D}'_S\mathbf{x}_S^t)(\mathbf{x}_S^{t'}\mathbf{y})^2 \\
&\quad\leq \sum_{t\in\mathcal{T}_n} w_t^\tau |g'(x_u^t\langle\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D},\mathbf{x}_S^t\rangle)(x_u^t\mathbf{D}'_S\mathbf{x}_S^t)|(\mathbf{x}_S^{t'}\mathbf{y})^2 \\
&\quad\leq \sqrt{s}||\mathbf{D}||_2\|\sum_t w_t^\tau\mathbf{x}_S^t\mathbf{x}_S^{t'}\|_2 \qquad (\ |g'(\cdot)|\leq 1\ ) \\
&\quad\leq D_{\max}K\lambda_n s \leq \frac{C_{\min}}{2}.
\end{aligned}$$

The last inequality follows as long as $\lambda_n s\leq\frac{C_{\min}}{2D_{\max}K}$. We have shown that

$$\Lambda_{\min}(\sum_{t\in\mathcal{T}_n} w_t^\tau \eta(\mathbf{x}^t;\boldsymbol{\theta}_u^\tau+\alpha\mathbf{D})\mathbf{x}_S^t\mathbf{x}_S^{t'})\geq\frac{C_{\min}}{2},$$

with high probability.

Putting the bounds on the three terms together, we have

$$H(\mathbf{D})\ \geq\ (\lambda_n\sqrt{s})^2\left\{-\frac{1}{4}K+\frac{C_{\min}}{2}K^2-K\right\},$$

which is strictly positive for $K=\frac{5}{C_{\min}}$. For this choice of K, we have that $\lambda_n s\leq\frac{C_{\min}^2}{10D_{\max}}$, which holds under the conditions of Theorem 1 for n large enough. □

References

[1] M. Arbeitman, E. Furlong, F. Imam, E. Johnson, B. Null, B. Baker, M. Krasnow, M. Scott, R. Davis, and K. White. Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, 297:2270–2275, 2002.

[2] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[3] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *APPROX '08 / RANDOM '08: Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop, RANDOM 2008 on Approximation, Randomization and Combinatorial Optimization*, pages 343–356, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] E. H. Davidson. *Genomic Regulatory Systems*. Academic Press, 2001.

[5] Mathias Drton and Michael D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

[6] J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. In *Proceedings of the Twenty-fourth Conference on Uncertainty in AI (UAI)*, 2008.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[8] Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541, 2009.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Department of Statistics, Stanford University, Tech. Rep*, 2008.

[10] Jerome Friedman, Trevor Hastie, Holger Hofling, and Robert Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008.

[12] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, August 2007.

[13] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming (web page and software), 2008.

[14] F. Guo, S. Hanneke, W. Fu, and E. P. Xing. Recovering temporally rewiring networks: A model-based approach. *International Conference of Machine Learning*, 2007.

[15] S. Hanneke and E.P Xing. Discrete temporal models of social networks. *Workshop on Statistical Network Analysis, the 23rd International Conference on Machine Learning (ICML-SNA)*, 2006.

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

[17] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519–1555, 2007.

[18] Mladen Kolar, Le Song, and Eric Xing. Sparsistent learning of varying-coefficient models with structural changes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1006–1014. 2009.

[19] S. L. Lauritzen. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, July 1996.

[20] N. Luscombe, M. Babu, H. Yu, M. Snyder, S. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

[21] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[22] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[23] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1 regularized logistic regression. *Annals of Statistics*, to appear, 2009.

[24] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Nov 2008.

[25] Alessandro Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5):2922–2952, 2009.

[26] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal Of Statistics*, 2:494, 2008.

[27] P. Sarkar and A. Moore. Dynamic social network analysis using latent space models. *Conference of Knowledge Discovery and Data Mining*, 2006.

[28] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal Of The Royal Statistical Society Series B*, 67(1):91–108, 2005.

[29] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

[30] M. A. J. van Duijn, K. J. Gile, and M. S. Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52 – 62, 2009.

[31] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

[32] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[33] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, March 2007.

[34] Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 455–466. Omnipress, 2008.

[35] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.