On Time Varying Undirected Graphs

Mladen Kolar
Machine Learning Department
Carnegie Mellon University

Eric P. Xing
Machine Learning Department
Carnegie Mellon University

Abstract

The time-varying multivariate Gaussian distribution and the undirected graph associated with it, as introduced in Zhou et al. (2008), provide a useful statistical framework for modeling complex dynamic networks. In many application domains, it is of high importance to estimate the graph structure of the model consistently for the purpose of scientific discovery. In this paper, we show that under suitable technical conditions, the structure of the undirected graphical model can be consistently estimated in the high dimensional setting, when the dimensionality of the model is allowed to diverge with the sample size. The model selection consistency is shown for the procedure proposed in Zhou et al. (2008) and for the modified neighborhood selection procedure of Meinshausen and Bühlmann (2006).

1 Introduction

Network models have become popular as a way to abstract complex systems and gain insights into relational patterns among observed variables. In many domains, including biology, astronomy and social sciences, particularly useful and successful network models are based on the Gaussian graphical models (GGMs). In the framework of the GGMs, the precision matrix, which is the inverse of the covariance matrix, represents conditional dependencies between random variables and a network representation is obtained by linking conditionally dependent variables. The hope is that this graphical representation is going to provide additional insight into the system under observation, for example, by showing how different parts of the system interact. A statistical challenge in this framework is to estimate reliably the precision matrix and the set of non-zero elements of the matrix from an observed sample.

Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ be a p-dimensional multivariate Gaussian random variable with mean zero and covariance $\mathbf{\Sigma}$. Let $\mathbf{\Omega} \triangleq \mathbf{\Sigma}^{-1}$ be the precision matrix. The (a, b)-element, ω_{ab}, of the precision matrix is proportional to the partial correlation between random variables X_a and X_b, the ath and bth component of $\mathbf{X}$. Therefore X_a is conditionally independent of X_b given the rest of variables if and only if $\omega_{ab} = 0$. This conditional dependence can be represented with a graph $G = (V, E)$, where the set of nodes V corresponds to the components of the random vector $\mathbf{X}$ and the edge set $E \subseteq V \times V$ includes edges between nodes only if the corresponding components are conditionally dependent, that is, an edge $e_{ab} \in E$ only if $\omega_{ab} \neq 0$. For a detailed account of the topic see, for example, Lauritzen (1996).

A large amount of literature in both statistics and machine learning has been devoted to the problem of estimating sparse precision matrices, where some elements are set to zero. The problem of estimating precision matrices with zeros is known in statistics as *covariance selection* and was introduced in the seminal paper by Dempster (1972). An introduction to classical approaches, which are commonly based on identifying the correct set of non-zero elements and then estimating the non-zero elements, can be found in, for example, Edwards (2000). These approaches are applicable only on data sets with a small number of variables and a large number of observations. However, due to the technological improvements of data collection processes, we have seen a surge in the number of high-dimensional data sets. As a result, more recent literature on estimating sparse precision matrices is focused on methods suitable for high-dimensional problems where the number of variables p can be much larger than the sample size n. A promising line of research, due to the scalability of algorithms and theo-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

retical guarantees of estimation procedures, estimates the precision matrix by minimizing a convex objective, which consists of a likelihood or a pseudo-likelihood term and a term accounting for the model complexity (see for example, Yuan and Lin, 2007; Fan et al., 2009; Banerjee et al., 2008; Rothman et al., 2008; Friedman et al., 2008; Ravikumar et al., 2008; Guo et al., 2010b; Zhou et al., 2008; Meinshausen and Bühlmann, 2006; Peng et al., 2009; Guo et al., 2010a; Wang et al., 2009). Due to the large number of investigations, the theory behind estimating sparse precision matrices is becoming settled.

While the most of the previous work deals with estimating a single precision matrix from *i.i.d.* samples and the static graph that it encodes, Zhou et al. (2008) studied the problem in which the probability distribution is allowed to vary with time. Formally, let

$$\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{t_i}), \quad i = 1, \ldots, n \tag{1}$$

be an independent sequence of p-dimensional observations distributed according to a multivariate normal distribution whose covariance matrix changes smoothly over time. Assume for simplicity that the time points are equidistant on a unit interval, that is, $t_i = i/n$. A graph $G^{t_i} = (V, E^{t_i})$ is associated with each observation $\mathbf{x}^i$ and it represents the non-zero elements of the precision matrix $\mathbf{\Omega}^{t_i} \triangleq (\mathbf{\Sigma}^{t_i})^{-1}$ (recall that $e_{ab} \in E^{t_i}$ only if $\omega_{ab}^{t_i} \neq 0$). With changing precision matrix $\mathbf{\Omega}^{t_i}$, the associated graphs change as well, which allows for modelling of dynamic networks. The model given in (1) can be thought of as a special case of the varying coefficient models introduced in Hastie and Tibshirani (1993). In particular, the model in (1), inherits flexibility and modelling power from the class of nonparametric models, but at the same time it retains interpretability of parametric models. Indeed, there are no assumptions on the parametric form of the elements of the covariance matrix $\mathbf{\Sigma}^t$ as a function of time.

Under the model (1), Zhou et al. (2008) studied the problem of the consistent recovery in the Frobenius norm of $\mathbf{\Omega}^\tau$ for some $\tau \in [0, 1]$, as well as the predictive performance of the fitted model. While those results are very interesting and important in statistics, in many application areas, it is the graph structure that provides most insight into complex systems by allowing visualiziation of relational structures and mechanisms that explain the data. For example, in computational biology, a graph estimated from a gene expression microarray profile can reveal the topology of genetic regulation circuitry, while in sociocultural analysis, a graph structure helps identify communities and communication patterns among actors. Unfortunately, the consistent estimation of the graph structure does not follow immediately from the consistent estimation of the precision matrix $\mathbf{\Omega}$. We address the problem of the consistent graph structure recovery under the model (1) in this paper. Our work has applications in many disciplines, including computational biology and computational finance, where the assumptions that the data are distributed *i.i.d.* are not satisfied. For example, a gene regulatory network is assumed to change throughout the developmental process of the organism, and a plausible way to model the longitudinal gene expression levels is by using the multivariate Gaussian distribution with a time-evolving structure.

The main contributions of the paper include establishing sufficient condition for the penalized likelihood procedure, proposed in Zhou et al. (2008), to estimate the graph structure consistently. Furthermore, we modify the neighborhood selection procedure of Meinshausen and Bühlmann (2006) to estimate the graph structure under the model (1) and provide sufficient conditions for the graph recovery.

1.1 Notation

The following notation is used throughout the paper. Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p \times p}$ be a matrix. Then, $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$, while $\varphi_{\min}(A)$ and $\varphi_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues, respectively. We use $\mathbf{A}^- \triangleq \mathbf{A} - \mathrm{diag}(\mathbf{A})$ to denote off-diagonal elements of $\mathbf{A}$. The ℓ_1 vector norm of the matrix $\mathbf{A}$ is given as $||\mathbf{A}||_1 \triangleq \sum_i \sum_j |a_{ij}|$. Similarly, we use the vector maximum norm $||\mathbf{A}||_\infty \triangleq \max_{i,j} |a_{ij}|$ to denote the element-wise maximum. The matrix Frobenius norm is denoted by $||\mathbf{A}||_F \triangleq \sqrt{\sum_i \sum_j a_{ij}^2}$. We will also use the (∞, ∞)-operator norm $|\!|\!|\mathbf{A}|\!|\!|_{\infty,\infty} \triangleq \max_i \sum_j |a_{ij}|$. Finally, we write $\mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{\mathrm{p}^2 \times 1}$ or $\vec{\mathbf{A}}$ for a vectorized form of matrix $\mathbf{A}$ obtained by stacking up the columns of $\mathbf{A}$. For a set $N \subset V$, we denote the set $\{X_a : a \in N\}$ as X_N. We use $\mathbf{X}$ to denote the $n \times p$ matrix whose rows consist of observations, with the vector $\mathbf{X}_a = (x_a^1, \ldots, x_a^n)'$ denoting a column a and, similarly, $\mathbf{X}_N = (\mathbf{X}_b : b \in N)$ denoting the $n \times |N|$ sub-matrix of $\mathbf{X}$ with columns indexed by the set N. For simplicity, we will use $\backslash a$ to denote the index set $\{1, \ldots, p\} \backslash \{a\}$, $\mathbf{X}_{\backslash a} = (\mathbf{X}_b : b \in \{1, \ldots, p\} \backslash \{a\})$. For a vector $\mathbf{a} \in \mathbb{R}^p$, we let $N(\mathbf{a})$ denote the set of non-zero components of $\mathbf{a}$.

2 Penalized likelihood estimation

In this section, we show that, under some technical conditions, the procedure proposed in Zhou et al. (2008) is able to consistently estimate the set of non-zero elements of the precision matrix $\mathbf{\Omega}^\tau$ at a given time point $\tau \in [0, 1]$. Under the model (1), an es-

timator of the precision matrix can be obtained by minimizing the following objective

$$\hat{\mathbf{\Omega}}^\tau = \underset{\mathbf{\Omega}\succ 0}{\operatorname{argmin}} \left\{ \operatorname{tr} \mathbf{\Omega}\hat{\mathbf{\Sigma}}^\tau - \log|\mathbf{\Omega}| + \lambda||\mathbf{\Omega}^-||_1 \right\}. \quad (2)$$

where $\hat{\mathbf{\Sigma}}^\tau = \sum_i w_i^\tau \mathbf{x}^i (\mathbf{x}^i)'$ is the weighted sample covariance matrix, with weights defined as

$$w_i^\tau = \frac{K_h(t_i - \tau)}{\sum_i K_h(t_i - \tau)}, \quad (3)$$

$K : \mathbb{R} \mapsto \mathbb{R}$ being the kernel function and $K_h(\cdot) = K(\cdot/h)$. The tuning parameter λ controls the number of non-zero pattern of the estimated precision matrix, while the bandwidth parameter h controls the smoothness over time of the estimated precision matrix and the effective sample size. These tuning parameters depend on the sample size n, but we will omit this dependence in our notation. In practice, the parameters are chosen using standard model selection techniques in data dependent way, for example, using cross-validation or Bayesian information criterion. The kernel K is taken such that the following set of assumptions holds.

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

For example, the assumption **K** is satisfied by the box kernel $K(x) = \frac{1}{2}\,\mathbb{1}\{x \in [-1, 1]\}$.

A similar estimator to the one given in (2) is analyzed in Zhou et al. (2008) and the convergence rate is established for $||\hat{\mathbf{\Omega}}^\tau - \mathbf{\Omega}^\tau||_F$. However, establishing that the estimated edge set

$$\hat{E}^\tau = \{(a, b) \ : \ a \neq b, \hat{\omega}_{ab}^\tau \neq 0\} \quad (4)$$

consistently estimates the true edge set $E^t = \{(a, b) : a \neq b, \omega_{ab}^t \neq 0\}$ is a harder problem, which requires stronger conditions on the true model. Let $s \triangleq \max_i |E^{t_i}|$ denote the maximum number of edges in a graph and $d \triangleq \max_i \max_{a\in V} |\{b \in V \mid a \neq b, e_{ab} \in E^{t_i}\}|$ the maximum node degree. In the remainder of this section, we provide sufficient conditions on (n, p, d, h, λ) under which the estimator given by (2) recovers the graph structure with high probability. To that end, we will use some of the results established in Ravikumar et al. (2008).

We start by imposing some assumptions on the true model. The first assumption assures that the covariance matrix is not singular at any time point. Note that if the population covariance matrix was singular, the problem of recovering the true graph structure would be ill-defined, since there would be no unique graph structure associated with the probability distribution.

Assumption C: There exist constants $\Lambda_{\max}, M_\infty < \infty$ such that for all $i \in \{1, \ldots, n\}$ we have

$$\frac{1}{\Lambda_{\max}} \leq \varphi_{\min}(\mathbf{\Sigma}^{t_i}) \leq \varphi_{\max}(\mathbf{\Sigma}^{t_i}) \leq \Lambda_{\max}$$

and

$$\left\|\mathbf{\Sigma}^{t_i}\right\|_{\infty,\infty} \leq M_\infty.$$

Furthermore, we assume that $\sigma_{ii}^\tau = 1$ for all $i \in V$.

The next assumption captures the notion of the distribution changing smoothly over time.

Assumption S: Let $\mathbf{\Sigma}^t = (\sigma_{ab}^t)$. There exists a constant $M_\Sigma > 0$ such that

$$\max_{a,b} \sup_{t\in[0,1]} |\dot{\sigma}_{ab}^t| \leq M_\Sigma, \text{ and}$$
$$\max_{a,b} \sup_{t\in[0,1]} |\ddot{\sigma}_{ab}^t| \leq M_\Sigma,$$

where $\dot{\sigma}_{ab}^t$ and $\ddot{\sigma}_{ab}^t$ denote the first and second derivative with respect to time.

Assumptions similar to **C** and **S** are also imposed in Zhou et al. (2008) in order to show consistency in the Frobenius norm. In particular, the rate of the convergence of $||\hat{\mathbf{\Omega}}^\tau - \mathbf{\Omega}^\tau||_F$ depends on the quantities $\Lambda_{\max}, M_\infty$ and M_Σ. Assumption **S** captures our notion of a distribution that is smoothly changing over time and together with assumption **C** guarantees that the precision matrix $\mathbf{\Omega}^t$ changes smoothly over time as well. The common variance of the components is assumed for presentation simplicity and can be obtained through scaling.

Assumptions **C** and **S** are not enough to guarantee recovery of the non-zero pattern of the population precision matrix $\mathbf{\Omega}^\tau$. From the previous work on variable selection in generalized linear models (see, for example, Fan and Lv (2009), Ravikumar et al. (2009), Bunea (2008)) we know that additional assumptions are needed on the Fisher information matrix in order to guarantee consistent model identification. In the case of the multivariate Gaussian distribution the Fisher information matrix at time $\tau \in [0, 1]$ is given as

$$\mathcal{I}^\tau \triangleq \mathcal{I}(\mathbf{\Omega}^\tau) = (\mathbf{\Omega}^\tau)^{-1} \otimes (\mathbf{\Omega}^\tau)^{-1},$$

where $\otimes$ denotes the Kronecker product. The elements of the Fisher information matrix can be also expressed as $\mathcal{I}^\tau_{(a,b),(a',b')} = \operatorname{Corr}(X_a^\tau X_b^\tau, X_{a'}^\tau X_{b'}^\tau)$. Let $S \triangleq S^\tau = E^\tau \cup \{(a, a)\}_{i\in V}$ be an index set of the non-zero elements of $\mathbf{\Omega}^\tau$ and S^C denotes its complement in $V \times V$. Let $\mathcal{I}^\tau_{SS}$ denote the $|S| \times |S|$ sub-matrix of $\mathcal{I}^\tau$ indexed by elements of S.

Assumption F: The sub-matrix $\mathcal{I}_{SS}$ is invertible. There exist constants $\alpha \in (0, 1]$ and $M_\mathcal{I} < \infty$ such that

$$\left\|\mathcal{I}^\tau_{S^C S}(\mathcal{I}^\tau_{SS})^{-1}\right\|_{\infty,\infty} \ \leq \ 1 - \alpha$$

and

$$\|(\mathcal{I}_{SS}^{\tau})^{-1}\|_{\infty,\infty} \;\leq\; M_{\mathcal{I}}.$$

The assumption **F** is identical to the assumptions made in Ravikumar et al. (2008). We need to assume that it holds only for the time point of interest τ at which the precision matrix is being estimated.

With these assumptions, we have the following result.

Theorem 1. *Fix a time point of interest $\tau \in [0,1]$. Let $\{\mathbf{x}^i\}$ be an independent sample according to the model (1). Under the assumptions* **C**, **S**, **F** *and* **K** *there exists a constant $C > 0$ depending only on $\Lambda_{\max}, M_\infty, M_\Sigma, M_K, M_{\mathcal{I}}$ and α for which the following holds. Suppose that the weighted sample covariance matrix $\hat{\mathbf{\Sigma}}^{\tau}$ is estimated using the kernel with the bandwidth parameter satisfying $h = \mathcal{O}\left(n^{-1/3}\right)$. If the penalty parameter λ in (2) scales as $\lambda = \mathcal{O}\left(n^{-1/3}\sqrt{\log p}\right)$ and the sample size satisfies $n > Cd^3(\log p)^{3/2}$, then the minimizer $\hat{\mathbf{\Omega}}^{\tau}$ of (2) defines the edge set $\hat{E}^{\tau}$ which satisfies*

$$\begin{aligned}\mathbb{P}[\hat{E}^{\tau} \neq \{(a,b) \;: a \neq b, |\omega_{ab}^{\tau}| > \omega_{\min}\}] \\ = \mathcal{O}(\exp(-c\log p)) \to 0,\end{aligned}$$

for some constant $c > 0$, with $\omega_{\min} = M_\omega n^{-1/3}\sqrt{\log p}$ and M_ω being a sufficiently large constant.

The theorem states that all the non-zero elements of the population precision matrix $\mathbf{\Omega}^{\tau}$, which are larger in absolute value than $\omega_{\min}$, will be identified. Note that if the elements of the precision matrix are too small, then the estimation procedure is not able to distinguish them from zero. Furthermore, the estimation procedure does not falsely include zero elements into the estimated set of edges. The theorem guarantees consistent recovery of the set of sufficiently large non-zero elements of the precision matrix at the time point τ. In order to obtain insight into the network dynamics, the graph corresponding to $\mathbf{\Omega}^t$ needs to be estimated at multiple time points. Due to the slow rate of convergence of $\hat{\mathbf{\Omega}}^t$, it is sufficient to estimate a graph at each time point t_i, $i = 1, \ldots, n$.

Comparing Theorem 1 to the results on the static graph structure estimation from an *i.i.d.* sample (Ravikumar et al., 2008), we can observe a slower rate of convergence. The difference arises from the fact that using the kernel estimate, we effectively use only the sample that is "close" to the time point τ. Using a local linear smoother, instead of the kernel smoother to reduce the bias in the estimation, a better dependence on the sample size could be obtained. Finally we note that, for simplicity and ease of interpretation, Theorem 1 is stated without providing explicit dependence of the rate of convergence on the constants appearing in the assumptions.

2.1 Proof of Theorem 1

The proof of the theorem will be separated into several propositions to facilitate the exposition. Technical lemmas and some proofs are given in the supplementary material. Our proof uses some ideas introduced in Ravikumar et al. (2008).

We start by introducing the following function

$$G(\mathbf{\Omega}) = \operatorname{tr}\mathbf{\Omega}\hat{\mathbf{\Sigma}}^{\tau} - \log|\mathbf{\Omega}| + \lambda||\mathbf{\Omega}^{-}||_1, \quad \forall \mathbf{\Omega} \succ 0$$

and we say that $\mathbf{\Omega} \in \mathbb{R}^{p\times p}$ satisfies the system $(\mathcal{S})$ when $\forall\; a \neq b \in V \times V$,

$$\begin{aligned}(\hat{\mathbf{\Sigma}}^{\tau})_{ab} - (\mathbf{\Omega}^{-1})_{ab} = -\lambda\,\mathrm{sign}((\mathbf{\Omega}^{-1})_{ab}), &\quad \text{if } (\mathbf{\Omega}^{-1})_{ab} \neq 0 \\ |(\hat{\mathbf{\Sigma}}^{\tau})_{ab} - (\mathbf{\Omega}^{-1})_{ab}| \leq \lambda, &\quad \text{if } (\mathbf{\Omega}^{-1})_{ab} = 0.\end{aligned} \tag{5}$$

It is known that $\mathbf{\Omega} \in \mathbb{R}^{p\times p}$ is the minimizer of Equation (2) if and only if it satisfies the system $(\mathcal{S})$. Since $G(\mathbf{\Omega})$ is strictly convex, the minimum, if attained, is unique. The assumption **C** guarantees that the minimum is attained. Therefore, we do not have to worry about the possibility of having several $\mathbf{\Omega}$ satisfying the system $(\mathcal{S})$.

Recall that we use the set S to index the non-zero elements of the population precision matrix. Without loss of generality we write

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{SS} & \mathcal{I}_{SS^C} \\ \mathcal{I}_{S^CS} & \mathcal{I}_{S^CS^C} \end{pmatrix}, \quad \vec{\mathbf{\Sigma}} = \begin{pmatrix} \vec{\mathbf{\Sigma}}_S \\ \vec{\mathbf{\Sigma}}_{S^C} \end{pmatrix}.$$

Let $\mathbf{\Omega} = \mathbf{\Omega}^{\tau} + \mathbf{\Delta}$. Using the first-order Taylor expansion of the function $g(\mathbf{X}) = \mathbf{X}^{-1}$ around $\mathbf{\Omega}^{\tau}$ we have

$$\mathbf{\Omega}^{-1} = (\mathbf{\Omega}^{\tau})^{-1} - (\mathbf{\Omega}^{\tau})^{-1}\mathbf{\Delta}(\mathbf{\Omega}^{\tau})^{-1} + R(\mathbf{\Delta}), \tag{6}$$

where $R(\mathbf{\Delta})$ denotes the remainder term. We consider the following two events

$$\begin{aligned}\mathcal{E}_1 \;=\; \Big\{|(\mathcal{I}_{SS})^{-1}[(\vec{\hat{\mathbf{\Sigma}}}^{\tau} - \vec{\mathbf{\Sigma}}^{\tau}) - \overrightarrow{R(\mathbf{\Delta})} + \\ \lambda\overrightarrow{\mathrm{sign}}(\mathbf{\Omega}^{\tau})]_S| < \omega(n,p)\Big\}\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{E}_2 = \Big\{|\mathcal{I}_{S^CS}(\mathcal{I}_{SS})^{-1}[(\vec{\mathbf{\Sigma}}^{\tau} - \vec{\hat{\mathbf{\Sigma}}}^{\tau}) + \overrightarrow{R(\mathbf{\Delta})}]_S + \\ (\vec{\hat{\mathbf{\Sigma}}}^{\tau} - \vec{\mathbf{\Sigma}}^{\tau})_{S^C} - \overrightarrow{R(\mathbf{\Delta})}_{S^C}| < \alpha\lambda\Big\},\end{aligned} \tag{8}$$

where, in both events, inequalities hold element-wise.

Proposition 2. *Under the assumptions of Theorem 1, the event*

$$\begin{aligned}\Big\{&\hat{\mathbf{\Omega}}^{\tau} \in \mathbb{R}^{p\times p} \text{ minimizer of (2)}, \\ &\mathrm{sign}(\hat{\omega}_{ab}) = \mathrm{sign}(\omega_{ab}^{\tau}) \text{ for all } |\omega_{ab}| \notin (0, \omega_{\min})\Big\}\end{aligned}$$

contains the event $\mathcal{E}_1 \cap \mathcal{E}_2$.

Proof. We start by manipulating the conditions given in (5). Using (6) and using the fact that $\text{vec}((\mathbf{\Omega}^\tau)^{-1}\mathbf{\Delta}(\mathbf{\Omega}^\tau)^{-1}) = ((\mathbf{\Omega}^\tau)^{-1} \otimes (\mathbf{\Omega}^\tau)^{-1})\vec{\mathbf{\Delta}} = \mathcal{I}\vec{\mathbf{\Delta}}$, we can rewrite (5) in the equivalent form

$$\begin{aligned}(\mathcal{I}\vec{\mathbf{\Delta}})_S + (\hat{\vec{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau)_S - (\overrightarrow{R(\mathbf{\Delta})})_S = -\lambda(\overrightarrow{\text{sign}}(\mathbf{\Omega}))_S \\ |(\mathcal{I}\vec{\mathbf{\Delta}})_{S^C} + (\hat{\vec{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau)_{S^C} - (\overrightarrow{R(\mathbf{\Delta})})_{S^C}| \leq \lambda \mathbb{1}_{S^C},\end{aligned} \tag{9}$$

where $\mathbb{1}_{S^C}$ is the vector of the form $(1, 1, \ldots, 1)'$ and the equations hold element-wise. Now consider the following linear functional, $F : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$,

$$\begin{aligned}\boldsymbol{\theta} \mapsto \boldsymbol{\theta} - \vec{\mathbf{\Omega}}^\tau_S + (\mathcal{I}_{SS})^{-1}\left[(\hat{\vec{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau) - \overrightarrow{R(\mathbf{\Delta})}\right]_S \\ + \lambda(\mathcal{I}_{SS})^{-1}\overrightarrow{\text{sign}}(\boldsymbol{\theta}).\end{aligned}$$

For any two vectors $\mathbf{x} = (x_1, \ldots, x_{|S|})' \in \mathbb{R}^{|S|}$ and $\mathbf{r} = (r_1, \ldots, r_{|S|})' \in \mathbb{R}^{|S|}_+$, define the set

$$\mathcal{B}(\mathbf{x}, \mathbf{r}) = \prod_{i=1}^{|S|}(x_i - r_i, x_i + r_i).$$

Now, we have $F(\mathcal{B}(\vec{\mathbf{\Omega}}^\tau_S, \omega_{\min})) =$

$$\begin{aligned}\mathcal{B}\big((\mathcal{I}_{SS})^{-1}[(\hat{\vec{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau) - \overrightarrow{R(\mathbf{\Delta})}]_S \\ + \lambda(\mathcal{I}_{SS})^{-1}\overrightarrow{\text{sign}}(\mathbf{\Omega}^\tau_S), \omega_{\min}\big) \triangleq \mathcal{H}.\end{aligned} \tag{10}$$

On the event $\mathcal{E}_1$, we have $\mathbf{0} \in \mathcal{H}$ and hence there exists $\hat{\vec{\mathbf{\Omega}}}_S \in \mathcal{B}(\vec{\mathbf{\Omega}}^\tau_S, \omega_{\min})$ such that $F(\hat{\vec{\mathbf{\Omega}}}_S) = \mathbf{0}$. Thus we have $\text{sign}(\bar{\omega}_{ab}) = \text{sign}(\omega^\tau_{ab})$ for all elements $(a, b) \in S$ such that $|\omega^\tau_{ab}| > \omega_{\min}$ and

$$\mathcal{I}_{SS}\vec{\mathbf{\Delta}}_S + (\hat{\vec{\mathbf{\Sigma}}} - \vec{\mathbf{\Sigma}})_S - (\overrightarrow{R(\mathbf{\Delta})})_S = -\lambda(\overrightarrow{\text{sign}}(\bar{\mathbf{\Omega}}))_S. \tag{11}$$

Under the assumption on the Fisher information matrix $\mathbf{F}$ and on the event $\mathcal{E}_2$ it holds

$$\begin{aligned}-\lambda \mathbb{1}_{S^C} &< \mathcal{I}_{S^CS}\vec{\mathbf{\Delta}}_S + \left(\hat{\vec{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau\right)_{S^C} - \left(\overrightarrow{R(\mathbf{\Delta})}\right)_{S^C} \\ &= \mathcal{I}_{S^CS}(\mathcal{I}_{SS})^{-1}\left[(\vec{\mathbf{\Sigma}}^\tau - \hat{\vec{\mathbf{\Sigma}}}^\tau) + \overrightarrow{R(\mathbf{\Delta})}\right]_S + \\ &\quad \left(\hat{\vec{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau\right)_{S^C} - \left(\overrightarrow{R(\mathbf{\Delta})}\right)_{S^C} + \\ &\quad \lambda\mathcal{I}_{S^CS}(\mathcal{I}_{SS})^{-1}(\overrightarrow{\text{sign}}\left(\bar{\mathbf{\Omega}}\right))_S \\ &< \lambda \mathbb{1}_{S^C}.\end{aligned} \tag{12}$$

Now, we consider the vector $\hat{\vec{\mathbf{\Omega}}} = \begin{pmatrix} \hat{\vec{\mathbf{\Omega}}}_S \\ \vec{\mathbf{0}}_{S^C} \end{pmatrix} \in \mathbb{R}^{p^2}$. Note that for $\bar{\mathbf{\Omega}}$, equations (11) and (12) are equivalent to saying that $\bar{\mathbf{\Omega}}$ satisfies conditions (9) or (5), that is, saying that $\bar{\mathbf{\Omega}}$ satisfies the system $(\mathcal{S})$. We have that $\text{sign}(\bar{\omega}_{ab}) = \text{sign}(\omega^\tau_{ab})$ for all (a, b) such that $|\omega^\tau_{ab}| \notin (0, \omega_{\min})$. Furthermore the solution to (2) is unique. □

Using Proposition 2, Theorem 1 follows if we show that events $\mathcal{E}_1$ and $\mathcal{E}_2$ occur with high probability. The following two propositions, with the proof also given in the supplementary materials, state that the events $\mathcal{E}_1$ and $\mathcal{E}_2$ occur with high probability.

Proposition 3. *Under the assumptions of Theorem 1, there exist constants $C_1, C_2 > 0$ depending on $\Lambda_{\max}$, M_∞, M_Σ, M_K, M_ω, $M_\mathcal{I}$ and α such that*

$$\mathbb{P}[\mathcal{E}_1] \geq 1 - C_1 \exp(-C_2 \log p).$$

Proposition 4. *Under the assumptions of Theorem 1, there exist $C_1, C_2 > 0$ depending on $\Lambda_{\max}$, M_∞, M_Σ, M_K, $M_\mathcal{I}$ and α such that*

$$\mathbb{P}[\mathcal{E}_2] \geq 1 - C_1 \exp(-C_2 \log p). \tag{13}$$

Now, Theorem 1 follows from Propositions 2, 3 and 4.

3 Neighborhood selection estimation

In this section, we discuss the neighborhood selection approach to selection of non-zero elements of the precision matrix $\mathbf{\Omega}^\tau$ under the model (1). The neighborhood selection procedure was proposed in Meinshausen and Bühlmann (2006) as a way to estimate the graph structure associated to a GGM from an *i.i.d.* sample. The method was applied to learn graph structure in more general settings as well (see, for example Ravikumar et al., 2009; Peng et al., 2009; Guo et al., 2010a; Kolar et al., 2010). As opposed to optimizing penalized likelihood, the neighborhood selection method is based on optimizing penalized pseudo-likelihood on each node of the graph, which results in local estimation of the graph structure. While the procedure is very scalable and suitable for large problems, it does not result in consistent estimation of the precision matrix. On the other hand, as we will show, the non-zero pattern of the elements of the precision matrix can be recovered under weaker assumptions.

We start by describing the neighborhood selection method under the model (1). As mentioned in the introduction, the elements of the precision matrix are related to the partial correlation coefficients as $\rho^t_{ab} = -\omega^t_{ab}/\sqrt{\omega^t_{aa}\omega^t_{bb}}$. A well known result (e.g., Lauritzen, 1996) relates the partial correlation coefficients to a regression model where a variable X_a is regressed onto the rest of variables $\mathbf{X}_{\setminus a}$,

$$X_a = \sum_{b \in V \setminus \{a\}} X_b \theta^t_{ab} + \epsilon^t_a, \quad a \in V. \tag{14}$$

In the equation above, ϵ^t_a is independent of $\mathbf{X}_{\setminus a}$ if and only if $\theta^t_{ab} = \rho^t_{ab}\sqrt{\omega^t_{aa}/\omega^t_{bb}}$. The relationship between the elements of the precision matrix and the

least square regression immediately suggests the following estimator for $\boldsymbol{\theta}^\tau_{\setminus a} \triangleq \{\theta^\tau_{ab}\}_{b\in V\setminus\{a\}}$,

$$\hat{\boldsymbol{\theta}}^\tau_{\setminus a} \triangleq \underset{\boldsymbol{\theta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \sum_i (x^i_a - \sum_{b\neq a} x^i_b\theta_b)^2 w^\tau_i + \lambda||\boldsymbol{\theta}||_1, \quad (15)$$

where the weight w^τ_i are defined in (3). The estimator $\hat{\boldsymbol{\theta}}^\tau_{\setminus a}$ defines the neighborhood of the node $a \in V$ at the time point τ as $\hat{N}^\tau_a \triangleq N(\hat{\boldsymbol{\theta}}^\tau_{\setminus a})$. By estimating the neighborhood of each node and combining them, the whole graph structure can be obtained. There are two natural ways to combine the estimated neighborhoods, using the union, $\hat{E}^{\tau,\cup} \triangleq \{(a,b) \ : \ b \in N^\tau_a \vee a \in N^\tau_b\}$, or intersection of different neighborhoods, $\hat{E}^{\tau,\cap} \triangleq \{(a,b) \ : \ b \in N^\tau_a \wedge a \in N^\tau_b\}$. Asymptotically these two approaches are equivalent and we will denote the resulting set of edges as $\hat{E}^\tau$.

The consistency of the graph estimation for the neighborhood selection procedure will be proven under similar assumptions to those of Theorem 1. However, the assumption **F** can be relaxed. Let $N \triangleq N^\tau_a \triangleq N(\boldsymbol{\theta}^\tau_{\setminus a})$ denote the set of neighbors of the node a. Using the index set N, we write $\boldsymbol{\Sigma}^\tau_{NN}$ for the $|N| \times |N|$ submatrix of $\boldsymbol{\Sigma}^\tau$ whose rows and columns are indexed by the elements of N.

Assumption $\tilde{\boldsymbol{F}}$: There exist constants $\gamma \in (0,1]$ such that

$$\|\boldsymbol{\Sigma}^\tau_{N^cN}(\boldsymbol{\Sigma}^\tau_{NN})^{-1}\|_{\infty,\infty} \ \leq \ 1-\gamma$$

for all $a = \{1,\ldots,p\}$ (recall that $N = N^\tau_a$).

The assumption $\tilde{\boldsymbol{F}}$ is known in the literature as the irrepresentable condition (van de Geer and Bühlmann, 2009; Wainwright, 2009; Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006). It is known that it is sufficient and almost necessary condition for the consistent variable selection in the Lasso setting. Compared to the assumption **F** that was sufficient for the consistent graph selection using penalized maximum likelihood estimator, the assumption $\tilde{\boldsymbol{F}}$ is weaker, see for example, Meinshausen (2008) and Ravikumar et al. (2008).

With these assumptions, we have the following result.

Theorem 5. *Fix a time point of interest $\tau \in [0,1]$. Let $\{\mathbf{x}^i\}$ be an independent sample according to the model (1). Under the assumptions* **C**, **S**, $\tilde{\boldsymbol{F}}$ *and* **K** *there exists a constant $C > 0$ depending only on $\Lambda_{\max}, M_\Sigma, M_K$ and γ for which the following holds. Suppose that the bandwidth parameter used in (15) satisfies $h = \mathcal{O}\left(n^{-1/3}\right)$. If the penalty parameter λ in (15) scales as $\lambda = \mathcal{O}\left(n^{-1/3}\sqrt{\log p}\right)$ and the sample size satisfies $n > Cd^{3/2}(\log p)^{3/2}$, then the neighborhood selection procedure defines the edge set $\hat{E}^\tau$, by solving (15) for all $a \in V$, which satisfies*

$$\mathbb{P}[\hat{E}^\tau \neq \{(a,b) \ : a \neq b, |\theta^\tau_{ab}| > \theta_{\min}\}] = \mathcal{O}(\exp(-cn^{2/3}(d\log p)^{-1})) \to 0,$$

for some constant $c > 0$, with $\theta_{\min} = M_\theta n^{-1/3}\sqrt{d\log p}$ and M_θ being a sufficiently large constant.

The theorem states that the neighborhood selection procedure can be used to estimate the pattern of nonzero elements of the matrix $\boldsymbol{\Omega}^\tau$ that are sufficiently large, as defined by $\theta_{\min}$ and the relationship between $\boldsymbol{\theta}^\tau_{\setminus a}$ and the elements of $\boldsymbol{\Omega}^\tau$. Similarly to the procedure defined in §2, in order to gain insight into the network dynamics, the graph structure needs to be estimated at multiple time points.

The advantage of the neighborhood selection procedure over the penalized likelihood procedure is that it allows for very simple parallel implementation, since the neighborhood of each node can be estimated independently. Furthermore, the assumptions under which the neighborhood selection procedure consistently estimates the structure of the graph are weaker. Therefore, since the network structure is important in many problems, it seems that the neighborhood selection procedure should be the method of choice. However, in problems where the estimated coefficients of the precision matrix are also of importance, the penalized likelihood approach has the advantage over the neighborhood selection procedure. In order to estimate the precision matrix using the neighborhood selection, one needs first to estimate the structure and then fit the parameters subject to the structural constraints. However, it was pointed out by Breiman (1996) that such two step procedures are not stable.

3.1 Proof of Theorem 5

There has been a lot of work on the analysis of the Lasso and related procedure (see for example Zhao and Yu (2006); Wainwright (2009); Bunea (2008); Bertin and Lecué (2008)). We will adapt some of the standard tools to prove our theorem. We will prove that the estimator $\hat{\boldsymbol{\theta}}^\tau_{\setminus a}$ defined in (15) consistently defines the neighborhood of the node a. Using the union bound over all the nodes in the graph, we will then conclude the theorem.

Unlike the optimization problem (2), the problem defined in (15) is not strongly convex. Let $\hat{\Theta}$ be the set of all minimizers of (15). To simplify the notation, we introduce $\tilde{\mathbf{X}}_a \in \mathbb{R}^{p-1}$ with components $\tilde{x}^i_a = \sqrt{w^\tau_i}x^i_a$ and $\tilde{\mathbf{X}}_{\setminus a} \in \mathbb{R}^{n\times p-1}$ with rows equal to $\tilde{\mathbf{x}}^i_{\setminus a} = \sqrt{w^\tau_i}\mathbf{x}^i_{\setminus a}$. With this, we say that $\boldsymbol{\theta} \in \mathbb{R}^{p-1}$ satisfies the system

$(\mathcal{T})$ when for all $b = 1, \ldots, p-1$

$$\begin{aligned} 2\tilde{\mathbf{X}}_b'(\tilde{\mathbf{X}}_a - \tilde{\mathbf{X}}_{\setminus a}\boldsymbol{\theta}) &= -\lambda \operatorname{sign}(\theta_b) \quad \text{if } \theta_b \neq 0 \\ |2\tilde{\mathbf{X}}_b'(\tilde{\mathbf{X}}_a - \tilde{\mathbf{X}}_{\setminus a}\boldsymbol{\theta})| &\leq \lambda \quad \text{if } \theta_b = 0. \end{aligned} \tag{16}$$

Furthermore, $\boldsymbol{\theta} \in \hat{\Theta}$ if and only if $\boldsymbol{\theta}$ satisfies the system $(\mathcal{T})$. The following result from Bunea (2008) relates the two elements of $\hat{\Theta}$.

Lemma 6. *Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be any two elements of $\hat{\Theta}$. Then $\tilde{\mathbf{X}}_{\setminus a}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) = 0$. Furthermore, all solutions have non-zero components in the same position.*

Proof. See Proposition 4.2 in Bunea (2008). □

The above lemma guarantees that even though the problem (15) is not strongly convex, all the solutions will define the same neighborhood.

Recall that $N = N_a$ denotes the set of neighbors of the node a. Without loss of generality, we can write

$$\hat{\boldsymbol{\Sigma}}^\tau = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}^\tau_{NN} & \hat{\boldsymbol{\Sigma}}^\tau_{NN^C} \\ \hat{\boldsymbol{\Sigma}}^\tau_{N^C N} & \hat{\boldsymbol{\Sigma}}^\tau_{N^C N^C} \end{pmatrix}.$$

We will consider the following two events

$$\mathcal{E}_3 = \left\{ |(2\hat{\boldsymbol{\Sigma}}^\tau_{NN})^{-1}[2\tilde{\mathbf{X}}_N'\mathbf{E} - \lambda \operatorname{sign}(\boldsymbol{\theta}^\tau_N)]| < \theta_{\min} \right\} \tag{17}$$

$$\mathcal{E}_4 = \Big\{ |2\hat{\boldsymbol{\Sigma}}^\tau_{N^C N}(\hat{\boldsymbol{\Sigma}}^\tau_{NN})^{-1}[\tilde{\mathbf{X}}_N'\mathbf{E} - \lambda \operatorname{sign}(\boldsymbol{\theta}^\tau_N)] - 2\tilde{\mathbf{X}}_{N^C}'\mathbf{E}| < \lambda \Big\}, \tag{18}$$

where, in both events, inequalities hold element-wise and $\mathbf{E} \in \mathbb{R}^n$ is the noise term with elements $e^i = \sqrt{w_i^\tau}(\epsilon_a^i + (\boldsymbol{\theta}^i_{\setminus a} - \boldsymbol{\theta}^\tau_{\setminus a})'\mathbf{x}^i)$. Note that the noise term is not centered and includes the bias term. In the light of Lemma 13, given in the supplementary material, the matrix $\hat{\boldsymbol{\Sigma}}^\tau_{NN}$ is invertible and the events $\mathcal{E}_3$ and $\mathcal{E}_4$ are well defined.

We have an equivalent of proposition 2 for the neighborhood selection procedure.

Proposition 7. *Under the assumptions of Theorem 5, the event*

$$\Big\{ \hat{\boldsymbol{\theta}}^\tau_{\setminus a} \in \mathbb{R}^{p-1} \text{ minimizer of (15)}, \\ \operatorname{sign}(\hat{\theta}_{ab}) = \operatorname{sign}(\theta^\tau_{ab}) \text{ for all } |\theta_{ab}| \notin (0, \theta_{\min}) \Big\}$$

contains the event $\mathcal{E}_3 \cap \mathcal{E}_4$.

The theorem 5 will follow from Proposition 7, once we show that the event $\mathcal{E}_3 \cap \mathcal{E}_4$ occurs with high-probability. The proof of Proposition 7 is based on the analysis of the conditions given in (16) and, since it follows the same reasoning given in the proof of Proposition 2, the proof is omitted.

The following two lemmas establish that the events $\mathcal{E}_3$ and $\mathcal{E}_4$ occur with high probability under the assumptions of Theorem 5.

Lemma 8. *Under the assumptions of Theorem 5, we have that*

$$\mathbb{P}[\mathcal{E}_3] \geq 1 - C_1 \exp(-C_2 \frac{nh}{d^2 \log d})$$

with constants C_1 and C_2 depending only on M_K, M_Σ, M_θ and $\Lambda_{\max}$.

Proof. To prove the lemma, we will analyze the following three terms separately,

$$T_1 = \lambda(2\hat{\boldsymbol{\Sigma}}^\tau_{NN})^{-1} \operatorname{sign}(\boldsymbol{\theta}^\tau_{\setminus a}), \tag{19}$$

$$T_2 = (2\hat{\boldsymbol{\Sigma}}^\tau_{NN})^{-1} 2\tilde{\mathbf{X}}_N'\mathbf{E}^1 \tag{20}$$

and

$$T_3 = (2\hat{\boldsymbol{\Sigma}}^\tau_{NN})^{-1} 2\tilde{\mathbf{X}}_N'\mathbf{E}^2, \tag{21}$$

where $\mathbf{E} = \mathbf{E}^1 + \mathbf{E}^2$, $\mathbf{E}^1 \in \mathbb{R}^n$ has elements $e^{i,1} = \sqrt{w_i^\tau}\epsilon_a^i$ and $\mathbf{E}^2 \in \mathbb{R}^n$ has elements $e^{i,2} = \sqrt{w_i^\tau}(\boldsymbol{\theta}^i_{\setminus a} - \boldsymbol{\theta}^\tau_{\setminus a})'\mathbf{x}^i$. Using the above defined terms and the triangle inequality, we need to show that $|T_1 + T_2 + T_3| \leq |T_1| + |T_2| + |T_3| < \theta_{\min}$.

Using Lemma 13, given in supplementary materials, we have the following chain of inequalities

$$\begin{aligned} ||T_1||_\infty \leq ||T_1||_2 &\leq 2\lambda \varphi_{\max}(\hat{\boldsymbol{\Sigma}}^{-1}_{NN})_2 || \operatorname{sign}(\boldsymbol{\theta}^\tau_{\setminus a})||_2 \\ &\leq C_1 \lambda \sqrt{d} \end{aligned}$$

with probability at least $1 - C_2 \exp(-C_3 \frac{nh}{d^2 \log d})$ and C_1, C_2 and C_3 are some constants depending on M_K and $\Lambda_{\max}$.

Next, we turn to the analysis of T_2. Conditioning on $\mathbf{X}_N$ and using Lemma 13, we have that the components of T_2 are normally distributed with zero mean and variance bounded by $C_1(nh)^{-1}$, where C_1 depends on $M_K, \Lambda_{\max}$. Next, using Gaussian tail bounds, we have that

$$||T_2||_\infty \leq C_1 \sqrt{\frac{\log d}{nh}}$$

with probability at least $1 - C_2 \exp(-C_3 \frac{nh}{d^2 \log d})$, where C_1 is a constant depending on M_K, $\Lambda_{\max}$ and M_Σ.

For the term T_3, we have that

$$\begin{aligned} ||T_3||_\infty \leq ||T_3||_2 &\leq \varphi_{\max}((\hat{\boldsymbol{\Sigma}}^\tau_{NN})^{-1}) ||\mathbf{E}^2||_2 \\ &\leq 2\Lambda_{\max} ||\mathbf{E}^2||_2 \end{aligned}$$

where the last inequality follows from an application of Lemma 13 with probability at least $1 - C_2 \exp(-C_3 \frac{nh}{d^2 \log d})$. Furthermore, elements of $\mathbf{E}^2$ are normally distributed with zero mean and variance

$C_1 hn^{-1}$. Hence, we can conclude that the term T_3 is asymptotically dominated by T_2.

Combining all the terms, we have that $|T_1 + T_2 + T_3| \leq M_\theta \frac{\sqrt{d \log p}}{n^{1/3}} = \theta_{\min}$ with probability at least $1 - C_1 \exp(-C_2 \frac{nh}{d^2 \log d})$ for constants C_1, C_2 and sufficiently large M_θ. □

Lemma 9. *Under the assumptions of Theorem 5, we have that*

$$\mathbb{P}[\mathcal{E}_4] \geq 1 - C_1 \exp(-C_2 \frac{nh}{d \log p})$$

with constants C_1 and C_2 depending only on $M_K, M_\Sigma, \Lambda_{\max}$ and γ.

Due to space constraints, the proof of the lemma is provided in the supplementary material.

Now, Theorem 5 follows from Propositions 7, 8 and 9 and an application of the union bound.

4 Discussion

In this paper, we focus on consistent estimation of the graph structure in high-dimensional time-varying multivariate Gaussian distributions, as introduced in Zhou et al. (2008). The non-parametric estimate of the sample covariance matrix used together with the ℓ_1 penalized log-likelihood estimation produces a good estimate of the concentration matrix. Our contribution is the derivation of the sufficient conditions under which the estimate consistently recovers the graph structure.

This work complements the earlier work on value consistent estimation of time-varying Gaussian graphical models in Zhou et al. (2008) in that the main focus here is the consistent structure recovery of the graph associated with the probability distribution at a fixed time point. Obtaining an estimator that consistently recovers the structure is a harder problem than obtaining an estimator that is only consistent in, say, Frobenius norm. However, the price for the correct model identification comes in much more strict assumptions on the underlying model. Note that we needed to assume the "irrepresentable-like" condition on the Fisher information matrix (Assumption **F**), which is not needed in the work of Zhou et al. (2008). In some problems, where we want to learn about the nature of the process that generates the data, estimating the structure of the graph associated with the distribution gives more insight into the nature than the values of the concentration matrix. This is especially true in cases where the estimated graph is sparse and easily interpretable by domain experts.

Motivated by many real world problems coming from diverse areas such as biology and finance, we extend the work of Ravikumar et al. (2008) which facilitates estimation under the assumption that the underlying distribution does not change. We assume that the distribution changes smoothly, an assumption that is more valid, but could still be unrealistic in real life. An interesting extension to this work would be to allow for abrupt changes in the distribution and the graph structure. There has been a lot of work done on estimating change points in the high-dimensional setting, see, for example, recent paper Harchaoui et al. (2009), and it would be interesting to incorporate a change point estimation into the framework presented here. Throughout the paper we have also assumed that the data is independent, but it is important to extend the theory to allow for dependent observations. This would allow for analysis of time series data, where it is often assumed that data is coming from a stationary process.

Furthermore, we extend the neighborhood selection procedure as introduced in Meinshausen and Bühlmann (2006) to the time-varying Gaussian graphical models. This is done in a straightforward way using ideas from the literature on the varying-coefficient models, where a kernel smoother is used to estimate the model parameters that change over time in an unspecified way. We have shown that the neighborhood selection procedure is a good alternative to the penalized log-likelihood estimation procedure, as it requires less strict assumptions on the model. In particular, the assumption **F** can be relaxed to $\tilde{\boldsymbol{F}}$. We believe that our work provides important insights into the problem of estimating structure of dynamic networks.

Acknowledgements

We would like to thank Larry Wasserman for many useful discussions and suggestions. The research reported here was supported in part by Grant ONR N000140910758, NSF DBI-0640543, NSF IIS-0713379 and a graduate fellowship from Facebook to MK.

References

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. ISSN 1533-7928.

Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241, 2008. doi: 10.1214/08-EJS327.

P.J. Bickel and E. Levina. Some theory for Fisher's linear discriminant function,'naive Bayes', and some

alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Leo Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996. ISSN 00905364.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153, 2008.

A. P. Dempster. Covariance selection. *Biometrics*, 28 (1):157–175, 1972. ISSN 0006341X.

David Edwards. *Introduction to Graphical Modelling*. Springer, June 2000. ISBN 0387950540.

J. Fan and J. Lv. Non-Concave Penalized Likelihood with NP-Dimensionality. *ArXiv e-prints*, October 2009.

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. doi: 10.1214/08-AOAS215.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint Structure Estimation for Categorical Markov Networks. 2010a.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint Estimation of Multiple Graphical Models. 2010b.

Zaïd Harchaoui, Francis Bach, and Éric Moulines. Kernel change-point analysis. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*. 2009.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. ISSN 00359246.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P. Xing. Estimating Time-Varying networks. *Annals of Applied Statistics*, 4(1):94—123, 2010.

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

N. Meinshausen. A note on the Lasso for graphical Gaussian model selection. *Statistics and Probability Letters*, 78(7):880–884, 2008.

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

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. doi: 10.1198/jasa.2009.0126.

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

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.

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

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, July 1990. ISBN 0126702306.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. 2009. ISBN 9780387790510.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. doi: 10.1214/09-EJS506.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009. ISSN 0018-9448. doi: 10.1109/TIT.2009.2016018.

Pei Wang, Dennis L Chao, and Li Hsu. Learning networks from high dimensional binary data: An application to genomic instability data. *0908.3882*, August 2009.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1): 19–35, March 2007. doi: 10.1093/biomet/asm018.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. ISSN 1533-7928.

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.

Supplementary material

We will use $C_1, C_2, \ldots$ as generic positive constants whose values may change from line to line.

Technical results of Section 2.1

In this section of the appendix we collect proofs of Section 2.1 and some additional technical results.

Some deviation results

Let $\hat{\mathbf{\Sigma}}^\tau = (\hat{\sigma}^\tau_{ab})$ and $\mathbf{\Sigma}^\tau = (\sigma^\tau_{ab})$. To bound the element-wise deviation of the weighted sample covariance matrix $\hat{\mathbf{\Sigma}}^\tau$ from the population covariance matrix Σ^τ, we use the following decomposition

$$|\sum_i w_i^\tau x_a^i x_b^i - \sigma^\tau_{ab}| \leq |\hat{\sigma}^\tau_{ab} - \mathbb{E}\hat{\sigma}^\tau_{ab}| + |\mathbb{E}\hat{\sigma}^\tau_{ab} - \sigma^\tau_{ab}|. \quad (22)$$

Standard treatment of the expectation integrals gives us that $|\mathbb{E}\hat{\sigma}^\tau_{ab} - \sigma^\tau_{ab}| = \mathcal{O}(h)$, see for example Tsybakov (2009). The following Lemma characterizes the first term in Equation (22).

Lemma 10. *Let $\tau \in [0,1]$ be a fixed time point. Assume that Σ^τ satisfies the assumptions* **S** *and* **C** *and the kernel function satisfies the assumption* **K**. *Let $\{\mathbf{x}^i\}$ be an independent sample according to the model* (1). *Then*

$$\mathbb{P}[|\hat{\sigma}^\tau_{ab} - \mathbb{E}\hat{\sigma}^\tau_{ab}| > \epsilon) \leq C_1 \exp(-C_2 nh\epsilon^2), \quad |\epsilon| \leq \delta, \quad (23)$$

where C_1, C_2 and δ depend only on $\Lambda_{\max}$ and M_K.

Proof. The argument is quite standard. We use some ideas presented in Bickel and Levina (2004). Let us define $\tilde{x}^i_a = \frac{x^i_a}{\sqrt{\sigma^i_{aa}}}$ and $\tilde{x}^i_b = \frac{x^i_b}{\sqrt{\sigma^i_{bb}}}$. Note that $\tilde{x}^i_a, \tilde{x}^i_b \sim \mathcal{N}(0,1)$ and $\mathrm{Corr}(\tilde{x}^i_a, \tilde{x}^i_b) = \rho^i_{ab}$, where

$$\rho^i_{ab} = \frac{\sigma^i_{ab}}{\sqrt{\sigma^i_{aa}\sigma^i_{bb}}}.$$

Now we have

$$\begin{aligned}
&\mathbb{P}[|\hat{\sigma}^\tau_{ab} - \mathbb{E}\hat{\sigma}^\tau_{ab}| > \epsilon] \\
&= \mathbb{P}[|\sum_i \frac{2}{nh} K_h(i-\tau)(x^i_a x^i_b - \sigma^i_{ab})| > \epsilon] \\
&= \mathbb{P}[|\sum_i \frac{2}{nh} K_h(i-\tau)\sqrt{\sigma^i_{aa}\sigma^i_{bb}}(\tilde{x}^i_a \tilde{x}^i_b - \rho^i_{ab})| > \epsilon].
\end{aligned}$$

A simple calculation gives that

$$\begin{aligned}
\tilde{x}^i_a \tilde{x}^i_b - \rho^i_{ab} = \frac{1}{4}\big((\tilde{x}^i_a + \tilde{x}^i_b)^2 - 2(1+\rho^i_{ab}) \\
- (\tilde{x}^i_a - \tilde{x}^i_b)^2 - 2(1-\rho^i_{ab})\big),
\end{aligned}$$

which combined with the equation above and union bound gives

$$\begin{aligned}
&\mathbb{P}\left[|\hat{\sigma}^\tau_{ab} - \mathbb{E}\hat{\sigma}^\tau_{ab}| > \epsilon\right] \\
&\quad \leq 2\mathbb{P}\left[M_K \Lambda_{\max} \sum_i \frac{4}{nh}((Z^i)^2 - 1) \geq \epsilon\right], \quad (24)
\end{aligned}$$

where Z^i are independent $\mathcal{N}(0,1)$. The lemma follows from the standard results on the large deviation of χ^2 random variables. □

The bandwidth parameter needs to be chosen to balance the bias and variance in (22). If the bandwidth is chosen as $h = \mathcal{O}(n^{-1/3})$, the following result is straight forward.

Lemma 11. *Under the assumptions* **K**, **S** *and* **C**, *if the bandwidth parameter satisfies $h = \mathcal{O}(n^{-1/3})$, then*

$$\mathbb{P}[\max_{a,b} |\hat{\sigma}^\tau_{ab} - \sigma^\tau_{ab}| > \epsilon) \leq C_1 \exp(-C_2 n^{2/3}\epsilon^2 + \log p),$$

where C_1 and C_2 are constants depending only on M_K, M_Σ and $\Lambda_{\max}$.

Proof. The lemma follows from (22) by applying the union bound. □

Next, we directly apply Lemma 5 and Lemma 6 from Ravikumar et al. (2008) to obtain bounds on the deviation term $\mathbf{\Delta} = \hat{\mathbf{\Omega}}^\tau - \mathbf{\Omega}^\tau$ and the remainder term $R(\mathbf{\Delta})$.

Lemma 12. *Assume that the conditions of Theorem 1 are satisfied. There exist constants $C_1, C_2 > 0$ depending only on $\Lambda_{\max}$, M_∞, M_Σ, M_K, $M_\mathcal{I}$ and α such that with probability at least $1 - C_1\exp(-C_2 \log p)$, the following two statements hold:*

1. *There exists some $M_\Delta > 0$ depending on $\Lambda_{\max}$, M_∞, M_Σ, M_K, $M_\mathcal{I}$ and α such that $||\mathbf{\Delta}||_\infty \leq M_\Delta n^{-1/3}\sqrt{\log p}$.*
2. *Furthermore, element-wise maximum of the remainder term $R(\mathbf{\Delta})$ can be bounded $||R(\mathbf{\Delta})||_\infty \leq \frac{\alpha\lambda}{8}$.*

Proof. We perform the analysis on the event $\mathcal{A}$ defined in (27). Under the assumption of the lemma, we have that $n > Cd^3(\log p)^{3/2}$ and on the event $\mathcal{A}$,

$$||\hat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma}^\tau||_\infty + \lambda \leq M_\Delta \lambda \leq \frac{M_\Delta}{d}. \quad (25)$$

This implies that under the conditions of Lemma 6 and Lemma 5 in Ravikumar et al. (2008) are satisfied and we apply them to conclude the statement of the lemma. □

The following lemma gives us deviation of the minimum eigenvalue of the weighted empirical covariance matrix from the population quantity.

Lemma 13. *Let $\tau \in [0,1]$ be a fixed time point. Assume that Σ^τ satisfies the assumptions* **S** *and* **C** *and the kernel function satisfies the assumption* **K**. *Let $\{\mathbf{x}^i\}$ be an independent sample according to the model (1). Then*

$$\mathbb{P}[|\Lambda_{\min}(\hat{\mathbf{\Sigma}}^\tau_{NN}) - \Lambda_{\min}(\mathbf{\Sigma}^\tau_{NN})| > \epsilon) \leq C_1 \exp(-C_2 \frac{nh}{|N|^2}\epsilon^2 + C_3 \log |N|), \tag{26}$$

where C_1, C_2 and C_3 are constants that depend only on $\Lambda_{\max}$, M_Σ and M_K.

Proof. Using perturbation theory results (see for example Stewart and Sun (1990)), we have that

$$|\Lambda_{\min}(\hat{\mathbf{\Sigma}}^\tau_{NN}) - \Lambda_{\min}(\mathbf{\Sigma}^\tau_{NN})| \leq ||\hat{\mathbf{\Sigma}}^\tau_{NN} - \mathbf{\Sigma}^\tau_{NN}||_F \leq |N| \max_{a\in N, b\in N} |\hat{\sigma}^\tau_{ab} - \sigma^\tau_{ab}|.$$

But then using (22), Lemma 10 and the union bound, the result follows. □

Proof of Proposition 3

We will perform analysis on the event

$$\mathcal{A} = \left\{ ||\hat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma}^\tau||_\infty \leq \frac{\alpha\lambda}{8} \right\}. \tag{27}$$

Under the assumptions of the proposition, it follows from Lemma 11 that $\mathbb{P}[\mathcal{A}] \geq 1 - C_1 \exp(-C_2 \log p)$. Also, under the assumptions of the proposition, Lemma 12 can be applied to conclude that $R(\mathbf{\Delta}) \leq \frac{\alpha\lambda}{8}$. Let $e_j \in \mathbb{R}^{|S|}$ be a unit vector with 1 at position j and zeros elsewhere. On the event $\mathcal{A}$, it holds that

$$\begin{aligned}
&\max_{1\leq j\leq |S|} |e_j'(\mathcal{I}_{SS})^{-1}[(\vec{\hat{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau) - \overrightarrow{R(\mathbf{\Delta})} + \lambda\overrightarrow{\operatorname{sign}}(\mathbf{\Omega}^\tau)]_S| \\
&\quad \leq |||(\mathcal{I}_{SS})^{-1}|||_{\infty,\infty}\big(||(\vec{\hat{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau)_S||_\infty + ||\overrightarrow{R(\mathbf{\Delta})}_S||_\infty \\
&\qquad + \lambda||\overrightarrow{\operatorname{sign}}(\mathbf{\Omega}^\tau_S)||_\infty\big) \\
&\qquad \text{(using the Hölder's inequality)} \\
&\quad \leq M_{\mathcal{I}}\frac{4+\alpha}{4}\lambda \leq C\frac{\sqrt{\log p}}{n^{1/3}} \\
&\quad < \omega_{\min} = M_\omega \frac{\sqrt{\log p}}{n^{1/3}},
\end{aligned}$$

for a sufficiently large constant M_ω.

Proof of Proposition 4

We will work on the event $\mathcal{A}$ defined in (27). Under the assumptions of the proposition, Lemma 12 gives $R(\mathbf{\Delta}) \leq \frac{\alpha\lambda}{8}$. Let $e_j \in \mathbb{R}^{p^2-|S|}$ be a unit vector with 1 at position j and zeros elsewhere. On the event $\mathcal{A}$, it holds that

$$\begin{aligned}
&\max_{1\leq j\leq (p^2-|S|)} \Big|e_j'\big(\mathcal{I}_{S^CS}(\mathcal{I}_{SS})^{-1}[(\vec{\mathbf{\Sigma}}^\tau - \vec{\hat{\mathbf{\Sigma}}}^\tau) + \overrightarrow{R(\mathbf{\Delta})}]_S + \\
&\qquad (\vec{\hat{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau)_{S^C} - \overrightarrow{R(\mathbf{\Delta})}_{S^C}\big)\Big| \\
&\quad \leq |||\mathcal{I}_{S^CS}(\mathcal{I}_{SS})^{-1}|||_{\infty,\infty}\Big(||\vec{\mathbf{\Sigma}}^\tau - \vec{\hat{\mathbf{\Sigma}}}^\tau||_\infty + ||\overrightarrow{R(\mathbf{\Delta})}||_\infty\Big) + \\
&\qquad ||\vec{\hat{\mathbf{\Sigma}}}^\tau - \vec{\mathbf{\Sigma}}^\tau||_\infty + ||\overrightarrow{R(\mathbf{\Delta})}||_\infty \\
&\quad \leq (1-\alpha)\frac{\alpha\lambda}{4} + \frac{\alpha\lambda}{4} \\
&\quad \leq \alpha\lambda,
\end{aligned}$$

which concludes the proof.

Technical results of Section 3

In this subsection, we provide a proof of Lemma 9.

Proof of Lemma 9

Only a proof sketch is provided here. We analyze the event defined in (18) by splitting it into several terms. Observe that for $b \in N^C$, we can write

$$\begin{aligned}
x^i_b &= \mathbf{\Sigma}^\tau_{bN}(\mathbf{\Sigma}^\tau_{NN})^{-1}\mathbf{x}^i_N \\
&\quad + [\mathbf{\Sigma}^{t_i}_{bN}(\mathbf{\Sigma}^{t_i}_{NN})^{-1} - \mathbf{\Sigma}^\tau_{bN}(\mathbf{\Sigma}^\tau_{NN})^{-1}]'\mathbf{x}^i_N \\
&\quad + v^i_b
\end{aligned}$$

where $v^i_b \sim \mathcal{N}(0, (\sigma^i_b)^2)$ with $\sigma^i_b \leq 1$. Let us denote $\tilde{\mathbf{V}}_b \in \mathbb{R}^n$ the vector with components $\tilde{v}^i_b = \sqrt{w^\tau_i}v^i_b$. With this, we have the following decomposition of the components of the event $\mathcal{E}_4$. For all $b \in N^c$,

$$w_{b,1} = \mathbf{\Sigma}^\tau_{bN}(\mathbf{\Sigma}^\tau_{NN})^{-1}\lambda\operatorname{sign}(\boldsymbol{\theta}^\tau_N),$$

$$w_{b,2} = \tilde{\mathbf{V}}'_b\Big[(\tilde{\mathbf{X}}_N(\hat{\mathbf{\Sigma}}_{NN})^{-1}\lambda\operatorname{sign}(\boldsymbol{\theta}^\tau_N) + \Pi^\perp_{\tilde{\mathbf{X}}_N}(\mathbf{E}^1)\Big],$$

$$w_{b,3} = \tilde{\mathbf{V}}'_b\Pi^\perp_{\tilde{\mathbf{X}}_N}(\mathbf{E}^2)$$

and

$$w_{b,4} = \tilde{\mathbf{F}}'_b\Big[(\tilde{\mathbf{X}}_N(\hat{\mathbf{\Sigma}}_{NN})^{-1}\lambda\operatorname{sign}(\boldsymbol{\theta}^\tau_N) + \Pi^\perp_{\tilde{\mathbf{X}}_N}(\mathbf{E}^1 + \mathbf{E}^2)\Big],$$

where $\Pi^\perp_{\tilde{\mathbf{X}}_N}$ is the projection operator defined as $\mathbf{I}_p - \tilde{\mathbf{X}}_N(\tilde{\mathbf{X}}'_N\tilde{\mathbf{X}}_N)^{-1}\tilde{\mathbf{X}}'_N$, $\mathbf{E}^1$ and $\mathbf{E}^2$ are defined in the proof of Lemma 8 and we have introduced $\tilde{\mathbf{F}}_b \in \mathbb{R}^n$ as the vector with components $\tilde{f}^i_b = \sqrt{w^\tau_i}[\mathbf{\Sigma}^{t_i}_{bN}(\mathbf{\Sigma}^{t_i}_{NN})^{-1} - \mathbf{\Sigma}^\tau_{bN}(\mathbf{\Sigma}^\tau_{NN})^{-1}]'\mathbf{x}^i_N$. The lemma will follow using the triangle inequality if we show that

$$\max_{b\in N^C} |w_{b,1}| + |w_{b,2}| + |w_{b,3}| + |w_{b,4}| \leq \lambda.$$

Under the assumptions of the lemma, it holds that $\max_{b\in N^C}|w_{b,1}| < (1-\gamma)\lambda$.

Next, we deal with the term $w_{b,2}$. We observe that conditioning on $\mathbf{X}_S$, we have that $w_{b,2}$ is normally distributed with variance that can be bounded combining results of Lemma 13 from the supplementary material with the proof of Lemma 4 in Wainwright (2009). Next, we use the Gaussian tail bound to conclude that $\max_{b \in N^C} |w_{b,2}| < \gamma\lambda/2$ with probability at least $1 - \exp(-C_2 nh(d \log p)^{-1})$.

An upper bound on the term $w_{b,3}$ is obtained as follows $w_{b,3} \leq ||\tilde{\mathbf{V}}_b||_2 ||\Pi^{\perp}_{\tilde{\mathbf{X}}_N}(\mathbf{E}^2)||_2$ and then observing that the term is asymptotically dominated by the term $w_{b,2}$. Using similar reasoning, we also have that $w_{b,4}$ is asymptotically smaller than $w_{b,2}$.

Combining all the upper bounds, we obtain the desired result.