Estimating Networks With Jumps

Mladen Kolar **Eric P. Xing**

School of Computer Science
Carnegie Mellon University
{mladenk, epxing}@cs.cmu.edu

Abstract

Network models have been popular for modeling and representing complex relationships and dependencies between observed variables. A network can be explored by estimating the sparse precision matrix of the observed variables. This work consider a scenario where data comes from a dynamic stochastic process, so that a single static network model cannot adequately capture transient dependencies, such as, gene regulatory dependencies throughout a developmental cycle of an organism. It is assumed that the data can be partitioned into a number of blocks, so that one precision matrix fits data in each block, which allows for modeling and exploration of more general data sets. Without knowing the number of blocks or the boundaries of the partitions, an estimation procedure is developed that jointly estimates the partition boundaries and the coefficient of the sparse precision matrix on each block of the partition. Convergence rates of both the partition boundaries and the network structure are established.

1 Introduction

In recent years, we have witnessed fast advancement of data-acquisition techniques in many areas, including biological domains, engineering and social sciences. As a result, new statistical and machine learning techniques are needed to help us develop better understanding of complexities underlying large, noisy data sets. Networks have been commonly used to abstract noisy data and provide an insight into regularities and dependencies between observed variables. Recent popular techniques for modeling and exploring networks estimate the sparse precision matrix, which is the inverse of the covariance matrix, since the elements of the precision matrix represent the associations or conditional covariances between corresponding variables. Once the sparse precision matrix is estimated, the network is drawn by connecting variables whose corresponding elements of the precision matrix are non-zero. The problem of estimating the precision matrix with zeros is known in the statistical literature as covariance selection, as introduced in the seminal paper by [3].

Let $\mathcal{D} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ be an independent and identically distributed sample according to a p-dimensional multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the covariance matrix. Let $\mathbf{\Omega} := \mathbf{\Sigma}^{-1}$ denote the precision matrix, with elements (ω_{ab}), $1 \leq a, b \leq p$. In the case of the multivariate normal distribution, the precision matrix $\mathbf{\Omega}$ encodes the conditional independence structure of the distribution and the pattern of the zero elements in the precision matrix define the structure of the associated graph G. In particular, an edge (a, b) is element of the edge set if and only if the element of the precision matrix ω_{ab} is non-zero. Covariance selection deals exactly with the problem of estimating the non-zero elements of $\mathbf{\Omega}$ from the sample $\mathcal{D}$. Since the work of [3], there has been a lot of work on model selection and parameter estimation of the precision matrix in the GGM, which we do not plan summarize here. Recently proposed method estimate the sparse precision matrix by optimizing penalized likelihood [20, 4, 1, 16, 5, 14, 22] or through neighborhood selection [10, 12, 6, 19], where the structure of the graph is estimated by estimating the neighborhood of each node. Both of these approaches are suitable for high-dimensional problems, even when $p \gg n$, and can be efficiently implemented using scalable convex program solvers.

Most of the above mentioned work assumes that one network model is sufficient to describe the dependencies in the observed data. Often such assumptions are not justified. For example, when data consists of microarray measurements of the gene expression levels collected throughout the cell cycle or development of an organism, different genes are active during different stages. This suggests that different distributions and hence different networks should be used to describe dependencies between measured variables at different time intervals. In this paper we are going to tackle the problem of estimating the structure of the GGM when the structure is allowed to change over time. By assuming that the parameters of the precision matrix change with time, we obtain extra flexibility to model a larger class of distributions while still retaining the interpretability of the static GGM. In particular, as the coefficients of the precision matrix change over time, we also allow the structure of the underlying graph to change as well.

Let $\{\mathbf{x}^i\}_{i\in[n]} \in \mathbb{R}^p$ be a sequence of n independent observations (we use $[n]$ to denote the set $\{1,\ldots,n\}$) from a p-dimensional multivariate normal distribution, not necessarily the same for every observation. Let $\{\mathcal{B}^j\}_{j\in[B]}$ be a disjoint partitioning of the set $[n]$ where each block of the partition consists of consecutive elements, that is, $\mathcal{B}^j \cap \mathcal{B}^{j'} = \emptyset$ for $j \neq j'$ and $\bigcup_j \mathcal{B}^j = [n]$ and $\mathcal{B}^j = [T_{j-1} : T_j] := \{T_{j-1}, T_{j-1}+1, \ldots, T_j - 1\}$. Let $\mathcal{T} := \{T_0 = 1 < T_1 < \ldots < T_B = n+1\}$ denote the set of partition boundaries. We will consider the following model

$$\mathbf{x}^i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}^j), \qquad i \in \mathcal{B}^j, \tag{1}$$

so that observations indexed by elements in $\mathcal{B}^j$ are p-dimensional realizations of a multivariate normal distribution with zero mean and the covariance matrix $\mathbf{\Sigma}^j = (\sigma^j_{ab})_{a,b\in[p]}$. Let $\mathbf{\Omega}^j := (\mathbf{\Sigma}^j)^{-1}$ denote the precision matrix with elements $(\omega^j_{ab})_{a,b\in[p]}$. With the number of partitions, B, and the boundaries of partitions, $\mathcal{T}$, unknown, we study the problem of estimating the non-zero elements of the precision matrices $\{\mathbf{\Omega}^j\}_{j\in[B]}$ from the sample $\{\mathbf{x}^i\}_{i\in[n]}$. In particular, we study the problem where the coefficients are piece-wise constant functions of time. A scenario where the coefficients are smoothly varying functions of time has been considered in [22] for the GGM and in [7] and [8] for the discrete MRF.

If the partitions $\{\mathcal{B}^j\}_j$ were known, the problem would be trivially reduced to the setting analyzed in the previous work. Dealing with the unknown partitions, together with the structure estimation of the model, calls for new methods. We propose and analyze a method based on *time-coupled* neighborhood selection, where the regression coefficients are forced to stay similar across time using a fusion-type penalty and the sparsity of each neighborhood is obtained through the ℓ_1 penalty. Details of the approach are given in §2. The structural changes are commonly determined through hypothesis testing and a separate linear model is fit to each of the estimated segments. In our work, we use the penalized model selection approach to jointly estimate the partition boundaries and the model parameters.

2 Graph estimation via Temporal-Difference Lasso

In this section, we introduce our covariance selection procedure, which is based on the neighborhood selection using the fused-type penalty. We call the proposed procedure Temporal-Difference Lasso (*TD-Lasso*). We start by reviewing the neighborhood selection procedure, which has previously been used to estimate graphs in, for example, [12, 10, 13, 6]. First, we relate the elements of the precision matrix $\mathbf{\Omega}$ to a regression problem. Let the set S_a to denote the neighborhood of the node a. Denote $\bar{S}_a$ the closure of S_a, $\bar{S}_a := S_a \cup \{a\}$, and N_a the set of nodes not in the neighborhood of the node a, $N_a = [p]\backslash\bar{S}_a$. It holds that $X_a \perp X_{N_a} | X_{S_a}$. The neighborhood of the node a can be easily seen from the non-zero pattern of the elements in the precision matrix $\mathbf{\Omega}$, $S_a = \{b \in [p]\backslash\{a\} \, : \, \omega_{ab} \neq 0\}$. Furthermore, it is a well known result for Gaussian graphical models that we can write $X_a = \sum_{b\in S_a} X_b\theta^a_b + \epsilon$, where ϵ is independent of $X_{\backslash a}$ and $\theta^a_b = -\omega_{ab}/\omega_{aa}$. Therefore, the neighborhood of a node a, S_a, is equal to the set of non-zero coefficients of $\boldsymbol{\theta}^a$. This relationship was used in [10] to estimate the coefficients $\boldsymbol{\theta}^a$ using the Lasso.

In this paper, we build on the neighbourhood selection procedure to estimate the changing graph structure in model (1). We use S^j_a to denote the neighborhood of the node a on the block $\mathcal{B}^j$ and

N_a^j to denote nodes not in the neighborhood of the node a on the j-th block, $N_a^j = V \backslash S_a^j$. The set S_a is used to denote the union of all neighborhoods of the node a, $S_a = \cup_{j \in [B]} S_a^j$. Consider the following estimation procedure

$$\hat{\boldsymbol{\beta}}^a = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1 \times n}} \underbrace{\sum_{i \in [n]} \left(x_a^i - \sum_{b \in \backslash a} x_b^i \beta_{b,i} \right)^2}_{\mathcal{L}(\boldsymbol{\beta})} + \text{pen}_{\lambda_1, \lambda_2}(\boldsymbol{\beta}) \tag{2}$$

where the penalty is defined as

$$\text{pen}_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) := 2\lambda_1 \sum_{i=2}^{n} ||\boldsymbol{\beta}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i-1}||_2 + 2\lambda_2 \sum_{i=1}^{n} \sum_{b \in \backslash a} |\beta_{b,i}|. \tag{3}$$

The penalty term is constructed from two terms. The first term ensures that the solution is going to be piecewise constant for some partition of $[n]$ (possibly a trivial one). The first term can be seen as a sparsity inducing term in the temporal domain, since it penalizes the difference between the coefficients $\boldsymbol{\beta}_{\cdot,i}$ and $\boldsymbol{\beta}_{\cdot,i+1}$ at successive time-points. The second term results in estimates that have many zero coefficients at each block of the partition. The estimated set of partition boundaries $\hat{\mathcal{T}} = \{\hat{T}_0 = 1\} \cup \{\hat{T}_j \in [2:n] \ : \ \hat{\boldsymbol{\beta}}^a_{\cdot,\hat{T}_j} \neq \hat{\boldsymbol{\beta}}^a_{\cdot,\hat{T}_j - 1}\} \cup \{\hat{T}_{\hat{B}} = n+1\}$ contains indices of points at which a change is estimated, with $\hat{B}$ being an estimate of the number of blocks B. The estimated number of the block $\hat{B}$ is controlled through the user defined penalty parameter λ_1, while the sparsity of the neighborhood is controlled through the penalty parameter λ_2.

Based on the estimated set of partition boundaries $\hat{\mathcal{T}}$, we can define the neighborhood estimate of the node a for each estimated block. Let $\hat{\boldsymbol{\theta}}^{a,j} = \hat{\boldsymbol{\beta}}^a_{\cdot,i}, \forall i \in [\hat{T}_{j-1} : \hat{T}_j]$ be the estimated coefficient vector for the block $\hat{\mathcal{B}}^j = [\hat{T}_{j-1} : \hat{T}_j]$. Using the estimated vector $\hat{\boldsymbol{\theta}}^{a,j}$, we define the neighborhood estimate of the node a for the block $\hat{\mathcal{B}}^j$ as $\hat{S}_a^j := S(\hat{\boldsymbol{\theta}}^{a,j}) := \{b \in \backslash a \ : \ \hat{\theta}_b^{a,j} \neq 0\}$. Solving (2) for each node $a \in V$ gives us a neighborhood estimate for each node. Combining the neighborhood estimates we can obtain an estimate of the graph structure for each point $i \in [n]$.

The choice of the penalty term is motivated by the work on penalization using total variation [15, 9], which results in a piece-wise constant approximation of an unknown regression function. The fusion-penalty has also been applied in the context of multivariate linear regression [17], where the coefficients that are spatially close, are also biased to have similar values. As a result, nearby coefficients are fused to the same estimated value. Instead of penalizing the ℓ_1 norm on the difference between coefficients, we use the ℓ_2 norm in order to enforce that all the changes occur at the same point.

2.1 Numerical procedure

Finding a minimizer $\hat{\boldsymbol{\beta}}^a$ of (2) can be a computationally challenging task for an off-the-shelf convex optimization procedure. We propose two use an accelerated gradient method with a smoothing technique [11], which converges in $\mathcal{O}(1/\epsilon)$ iterations where ϵ is the desired accuracy.

We start by defining a smooth approximation of the fused penalty term. Let $\mathbf{H} \in \mathbb{R}^{n \times n-1}$ be a matrix with elements -1 when $i = j$, 1 when $i = j+1$ and 0 otherwise. With the matrix $\mathbf{H}$ we can define the following smooth approximation to the fused penalty

$$\Psi_\mu(\boldsymbol{\beta}) := \max_{\mathbf{U} \in \mathcal{Q}} \langle\langle \mathbf{U}, 2\lambda_1 \boldsymbol{\beta}\mathbf{H} \rangle\rangle - \mu ||\mathbf{U}||_F^2$$

where $\mathcal{Q} := \{\mathbf{U} \in \mathbb{R}^{p-1 \times n-1} \ : \ ||\mathbf{U}_{\cdot,i}||_2 \leq 1, \ \forall i \in [n-1]\}$ and $\mu > 0$ is the smoothness parameter. We have that $\Psi_\mu(\boldsymbol{\beta}) \leq \Psi_0(\boldsymbol{\beta}) \leq \Psi_\mu(\boldsymbol{\beta}) + \mu(n-1)$. Setting the smoothness parameter to $\mu = \frac{\epsilon}{2(n-1)}$, the correct rate of convergence is ensured. From [11], we have that $\Psi_\mu(\boldsymbol{\beta})$ is continuously differentiable and convex, with the gradient

$$\nabla \Psi_\mu(\boldsymbol{\beta}) = 2\lambda_1 \Pi_{\mathcal{Q}}(\frac{\lambda \boldsymbol{\beta}\mathbf{H}}{\mu})\mathbf{H}' \tag{4}$$

that is Lipschitz continuous. Here $\Pi_{\mathcal{Q}}(\cdot)$ is the projection operator onto the set $\mathcal{Q}$.

```
**Algorithm:** Accelerated Gradient Method for Equation (2)
**Input**: $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta}_0 \in \mathbb{R}^{p-1\times n}$, $\gamma > 1$, $L > 0$, $\mu = \frac{\epsilon}{2(n-1)}$
**Output**: $\hat{\boldsymbol{\beta}}^a$
Initialize $k := 1$, $\alpha_k := 1$, $\mathbf{z}_k := \boldsymbol{\beta}_0$
repeat
    while $F(p_L(\mathbf{z}_k)) > Q_L(p_L(\mathbf{z}_k), \mathbf{z}_k)$ do
        $L := \gamma L$
    $\boldsymbol{\beta}_k := p_L(\mathbf{z}_k)$        (using Eq. (6))
    $\alpha_{k+1} := \frac{1+\sqrt{1+4\alpha_k}}{2}$
    $\mathbf{z}_{k+1} := \boldsymbol{\beta}_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1})$
until *convergence*
$\hat{\boldsymbol{\beta}}^a := \boldsymbol{\beta}_k$
```

With the above defined smooth approximation, we focus on minimizing the following objective

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1\times n}} F(\boldsymbol{\beta}) := \min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1\times n}} \mathcal{L}(\boldsymbol{\beta}) + \Psi_\mu(\boldsymbol{\beta}) + 2\lambda_2||\boldsymbol{\beta}||_1.$$

Following [2], we define the following quadratic approximation of $F(\boldsymbol{\beta})$ at a point $\boldsymbol{\beta}_0$

$$Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}_0) := \mathcal{L}(\boldsymbol{\beta}_0) + \Psi_\mu(\boldsymbol{\beta}_0) + \langle\langle \boldsymbol{\beta} - \boldsymbol{\beta}_0, \nabla\mathcal{L}(\boldsymbol{\beta}_0) + \nabla\Psi(\boldsymbol{\beta}_0)\rangle\rangle + \frac{L}{2}||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_F^2 + 2\lambda_2||\boldsymbol{\beta}||_1 \quad (5)$$

where $L > 0$ is the parameter chosen as an upper bounds for the Lipschitz constant of $\nabla\mathcal{L} + \nabla\Psi$. Let $p_L(\boldsymbol{\beta}_0)$ be the minimizer of $Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$, which can be obtained in a closed form, as a result of the soft-thresholding,

$$p_L(\boldsymbol{\beta}_0) = T\left(\boldsymbol{\beta}_0 - \frac{1}{L}\big(\nabla\mathcal{L} + \nabla\Psi\big)(\boldsymbol{\beta}_0), \frac{2\lambda_2}{L}\right) \quad (6)$$

where $T(x, \lambda) = \mathrm{sign}(x)\max(0, |x| - \lambda)$ is the element-wise soft-thresholding operator.

In practice, an upper bound on the Lipschitz constant of $\nabla\mathcal{L} + \nabla\Psi$ can be expensive to compute, so the parameter L is going to be determined iteratively. The algorithm is given at the top of the page. The constant γ is used to increase the estimate of the Lipschitz constant L. From [2] we have that the algorithm converges in $\mathcal{O}(1/\epsilon)$ iterations.

3 Theoretical results

This section is going to address the statistical properties of the estimation procedure presented in Section 2. The properties are addressed in a double asymptotic framework where the dimensionality $p = p(n)$ is allowed to grow with the sample size. For the simplicity of the presentation, the number of blocks B is assumed to be fixed and known in advance[1] For the asymptotic framework to make sense, we assume that there exists a fixed unknown sequence of numbers $\{\tau_j\}$ that defines the partition boundaries as $T_j = \lfloor n\tau_j \rfloor$, where $\lfloor a \rfloor$ denotes the largest integer smaller that a. This assures that as the number of samples grow, the same fraction of samples falls into every partition. We call $\{\tau_j\}$ the boundary fractions.

We give sufficient conditions under which the sequence $\{\tau_j\}$ is consistently estimated. In particular, we show that $\max_{j\in[B]} |\hat{T}_j - T_j| \le n\delta_n$ with probability tending to 1, where $\{\delta_n\}_n$ is a non-increasing sequence of positive numbers that tends to zero. With the boundary segments consistently estimated, we further show that under suitable conditions for each node $a \in V$ the correct neighborhood is selected on all estimated block partitions that are sufficiently large.

[1] We have also a stronger result that does not need this assumption, but will not include it here.

The proof technique employed in this section is quite involved, so we briefly describe the steps used. Our analysis is based on careful inspection of the optimality conditions that a solution $\hat{\boldsymbol{\beta}}^a$ of the optimization problem (2) need to satisfy. The Karush-Kuhn-Tucker (KKT) conditions, which are sufficient and necessary for $\hat{\boldsymbol{\beta}}^a$ to be a solution of (2), are given in §3.2. Using the optimality conditions, we establish the rate of convergence for the partition boundaries. This is done by proof by contradiction. Suppose that there is a solution with the partition boundary $\hat{\mathcal{T}}$ that if far from $\mathcal{T}$. Then we show that, with high-probability, all such solutions will not satisfy the KKT conditions and therefore cannot be optimal. This shows that all the solutions to the optimization problem (2) result in partition boundaries that are "close" to the true partition boundaries, with high-probability. We can further show that the neighborhood estimates are consistently estimated, under the assumption that the estimated blocks of the partition have enough samples. Our analysis is going to focus on one node $a \in V$ and its neighborhood. However, using the union bound over all nodes in V, we will be able to carry over conclusions to the whole graph. To simplify our notation, when it is clear from the context, we will omit the superscript a and write $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$ and S, etc., to denote $\hat{\boldsymbol{\beta}}^a$, $\hat{\boldsymbol{\theta}}^a$ and S_a, etc.

3.1 Assumptions

Before presenting our theoretical results, we give some definitions and assumptions that are going to be used in this section. Let $\Delta_{\min} := \min_{j\in[B]} |T_j - T_{j-1}|$ denote the minimum length between change points, $\xi_{\min} := \min_{a\in V} \min_{j\in[B-1]} ||\boldsymbol{\theta}^{a,j+1} - \boldsymbol{\theta}^{a,j}||_2$ denote the minimum jump size and $\theta_{\min} = \min_{a\in V} \min_{j\in[B]} \min_{b\in S^j} |\theta_b^{a,j}|$ the minimum coefficient size. Throughout the section, we assume that the following holds.

A1 There exist two constants $\phi_{\min} > 0$ and $\phi_{\max} < \infty$ such that $\phi_{\min} = \min\{\Lambda_{\min}(\boldsymbol{\Sigma}^j_{S_aS_a}) : j \in [B], a \in V\}$ and $\phi_{\max} = \max\{\Lambda_{\max}(\boldsymbol{\Sigma}^j_{S_aS_a}) : j \in [B], a \in V\}$.

A2 Variables are scaled so that $\sigma^j_{aa} = 1$ for all $j \in [B]$ and all $a \in V$.

The assumption **A1** is commonly used to ensure that the model is identifiable. If the population covariance matrix is ill-conditioned, the question of the correct model identification if not well defined, as a neighborhood of a node may not be uniquely defined. The assumption **A2** is assumed for the simplicity of the presentation. The common variance can be obtained through scaling.

A3 There exists a constant $M > 0$ such that $\max_{a\in V} \max_{j,k\in[B]} \|\boldsymbol{\theta}^{a,k} - \boldsymbol{\theta}^{a,j}\|_2 \leq M$.

The assumption **A3** states that the difference between coefficients on two different blocks, $||\boldsymbol{\theta}^{a,k} - \boldsymbol{\theta}^{a,j}||_2$, is bounded for all $j, k \in [B]$. This assumption is simply satisfied if the coefficients $\boldsymbol{\theta}^a$ were bounded in the ℓ_2 norm.

A4 There exist a constant $\alpha \in (1/2, 1]$, such that the following holds

$$\sup_{\tilde{\boldsymbol{\Sigma}}} |||\tilde{\boldsymbol{\Sigma}}_{N_aS_a}(\tilde{\boldsymbol{\Sigma}}_{S_aS_a})^{-1}|||_\infty \leq 1 - \alpha, \qquad \forall a \in V,$$

where the supremum is taken over $\{\sum_j \gamma_j \boldsymbol{\Sigma}^j \ : \ \gamma_j \geq 0, \ \sum_j \gamma_j = 1\}$.

The assumption **A4** states that the variables in the neighborhood of the node a, S_a, are not too correlated with the variables in the set N_a. We need this assumption in order to prove that even when the partition boundaries are not correctly estimated, the neighborhood estimated on a block of the estimated partition does not include variables from the set N_a. This assumption is more restrictive than the commonly assumed irrepresentable condition (see for example [21, 18]), which is sufficient and necessary for the correct identification of the neighborhood on the block $\mathcal{B}^j$ if the partition boundaries were known.

A5 There exists a small constant $\delta > 0$, so that for all $\delta_1, \delta_2 \in [0, \delta)$ the following holds

$$|||\tilde{\boldsymbol{\Sigma}}^j_{N_a^jS_a^j}(\tilde{\boldsymbol{\Sigma}}^j_{S_a^jS_a^j})^{-1}|||_\infty \leq 1 - \alpha, \qquad \forall j \in [B], \forall a \in V,$$

where $\tilde{\boldsymbol{\Sigma}}^j = (1 + \delta_1 + \delta_2)^{-1}(\boldsymbol{\Sigma}^j + \delta_1\boldsymbol{\Sigma}^{j-1} + \delta_2\boldsymbol{\Sigma}^{j+1})$ and α is defined in **A4**.

The condition **A5** is again related to the correct estimation of the neighborhood on the block $\mathcal{B}^j$. The assumption is needed in order to show that even when the partition boundaries are not exactly recovered, the correct neighborhood can be estimated on the block $\mathcal{B}^j$.

A6 There exists a constant $\delta_p > 0$ such that the number of variables $p = p_n$ satisfy $p_n = \mathcal{O}(n^{\delta_p})$. The maximum degree of a node is assumed to be constant, $s = \mathcal{O}(1)$. The minimum coefficient size $\theta_{\min}$ satisfies $\theta_{\min} = \Omega(\sqrt{\log n/n})$.

The assumption implies that our procedure can be used to estimate the graph structure in a high-dimensional setting with $p \gg n$. The bound on the degree of nodes is imposed for the simplicity of the presentation and can be relaxed at the expense of more complex proofs. Our arguments can be modified to allow for the maximum degree of a node s to grow with the sample size. The lower bound on the minimum coefficient size $\theta_{\min}$ is necessary, since if a partial correlation coefficient is too close to zero the edge in the graph would not be detectable.

A7 The sequence of partition boundaries $\{T_j\}$ satisfy $T_j = \lfloor n\tau_j \rfloor$, where $\{\tau_j\}$ is a fixed, unknown sequence of the boundary fractions belonging to $[0, 1]$.

The assumption is needed for the asymptotic setting. As $n \to \infty$, there will be enough sample points in each of the blocks to estimate the neighborhood of nodes correctly.

3.2 Characterization of solutions

Although the optimization problem in (2) is convex, there may be multiple solutions to it, since it is not strictly convex. Using Karush-Kuhn-Tucker conditions, we can characterize any solution of (2).

Lemma 1. *A matrix $\hat{\boldsymbol{\beta}}$ is optimal for the optimization problem* (2) *if and only if there exist a collection of subgradient vectors $\{\hat{\mathbf{z}}^i\}_{i\in[2:n]}$ and $\{\hat{\mathbf{y}}^i\}_{i\in[n]}$, with $\hat{\mathbf{z}}^i \in \partial||\hat{\boldsymbol{\beta}}_{\cdot,i} - \hat{\boldsymbol{\beta}}_{\cdot,i-1}||_2$ and $\hat{\mathbf{y}}^i \in \partial||\hat{\boldsymbol{\beta}}_{\cdot,i}||_1$, that satisfies*

$$\sum_{i=k}^{n} \mathbf{x}^i_{\backslash a}\langle \mathbf{x}^i_{\backslash a}, \hat{\boldsymbol{\beta}}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i}\rangle - \sum_{i=k}^{n} \mathbf{x}^i_{\backslash a}\epsilon^i + \lambda_1 \hat{\mathbf{z}}^k + \lambda_2 \sum_{i=k}^{n} \hat{\mathbf{y}}^i = 0 \tag{7}$$

for all $k \in [n]$ and $\hat{\mathbf{z}}^1 = \hat{\mathbf{z}}^{n+1} = \mathbf{0}$.

While there may be multiple solutions to the problem (2), under some conditions, we can characterize the sparsity pattern of any solution that has specified partition boundaries $\hat{\mathcal{T}}$.

Lemma 2. *Let $\hat{\boldsymbol{\beta}}$ be a solution to* (2)*, with $\hat{\mathcal{T}}$ being an associated estimate of the partition boundaries. Suppose that the subgradient vectors satisfy $|\hat{y}^i_b| < 1$ for all $b \notin S(\hat{\boldsymbol{\beta}}_{\cdot,i})$, then any other solution $\tilde{\boldsymbol{\beta}}$ with the partition boundaries $\hat{\mathcal{T}}$ satisfy $\tilde{\beta}_{b,i} = 0$ for all $b \notin S(\hat{\boldsymbol{\beta}}_{\cdot,i})$.*

The above Lemma states sufficient conditions under which the sparsity patter of a solution with the partition boundary $\hat{\mathcal{T}}$ is unique. Note, however, that there may other solutions to (2) that have different partition boundaries.

3.3 Some results

Suppose that we know that there is a solution to the optimization problem (2) with the partition boundary $\hat{\mathcal{T}}$. Then that solution is also a minimizer of the following objective

$$\min_{\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{\hat{B}}} \sum_{j\in\hat{B}} ||\mathbf{X}_a^{\hat{\mathcal{B}}^j} - \mathbf{X}_{\backslash a}^{\hat{\mathcal{B}}^j}\boldsymbol{\theta}^j||_2^2 + 2\lambda_1 \sum_{j=2}^{\hat{B}} ||\boldsymbol{\theta}^j - \boldsymbol{\theta}^{j-1}||_2 + 2\lambda_1 \sum_{j=1}^{\hat{B}} |\hat{\mathcal{B}}^j|||\boldsymbol{\theta}^j||_1. \tag{8}$$

Note that the problem (8) does not give a practical way of solving (2), but will help us to reason about the solutions of (2). Let $\{r_n\}$ is an increasing sequence, to be characterized below. Our goal is to characterize the neighborhood of the node a whenever the estimated block of the partition contains more samples than r_n. We have the following proposition.

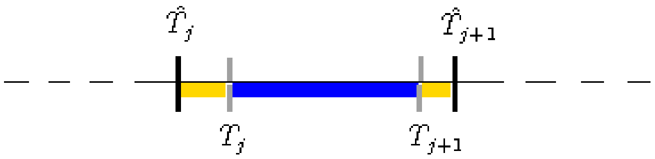

Figure 1: The figure illustrates where we expect to estimate a neighborhood of a node consistently. The blue region corresponds to the overlap between the true block (bounded by gray lines) and the estimated block (bounded by black lines). If the blue region is much larger than the orange regions, the additional bias introduced from the samples from the orange region will not considerably affect the estimation of the neighborhood of a node on the blue region. However, we cannot hope to consistently estimate the neighborhood of a node on the orange region.

Proposition 3. *Let $\{\hat{\beta}_{\cdot,i}\}_{i\in[n]}$ be any solution of* (2) *and let $\hat{\mathcal{T}}$ be the associated estimate of the block partition. Assume that* **A1**-**A4** *and* **A6**-**A7** *hold. Let $\{r_n\}_{n\geq 1}$ be an increasing sequence of numbers that satisfy $(r_n\lambda_2)^{-1}\lambda_1 \to 0$ and $r_n\lambda_2^2 \to \infty$ as $n \to \infty$. Then for all blocks $\hat{\mathcal{B}}^k$ defined by the partition $\hat{\mathcal{T}}$ that satisfy $|\hat{\mathcal{B}}^k| \geq r_n$ we have*

$$\mathbb{P}[S(\hat{\boldsymbol{\theta}}^k) \subseteq S] \to 1.$$

The above statement holds uniformly for all solutions $\{\hat{\beta}_{\cdot,i}\}_{i\in[n]}$ of (2).

Suppose that the penalty parameters satisfy

$$\lambda_1 \asymp \lambda_2 = \mathcal{O}(\sqrt{\log n/n}), \tag{9}$$

then, using proposition 3, on the blocks that are of size at least $r_n = \Omega(n/\log n)$, we have that the estimated neighborhood is contained in S. We will use this fact to prove the following result on the convergence rate of the estimated boundaries of $\hat{\mathcal{T}}$. First, under the assumption that the correct number of blocks is known.

Theorem 4. *Let $\{\mathbf{x}^i\}_{i\in[n]}$ be a sequence of observation according to the model in* (1). *Assume that the conditions of proposition 3 are satisfied and that the penalty parameters λ_1 and λ_2 satisfy* (9). *Let $\{\delta_n\}_{n\geq 1}$ be a non-increasing positive sequence that converges to zero as $n \to \infty$ and satisfies $\Delta_{\min} \geq n\delta_n$ for all $n \geq 1$. Furthermore, suppose that $(n\delta_n\xi_{\min})^{-1}\lambda_1 \to 0$, $\xi_{\min}^{-1}\sqrt{s}\lambda_2 \to 0$ and $(\xi_{\min}\sqrt{n\delta_n})^{-1}\sqrt{s\log n} \to 0$, then if $|\hat{\mathcal{T}}| = B + 1$ the following holds*

$$\mathbb{P}[\max_{j\in[B]} |T_j - \hat{T}_j| \leq n\delta_n] \to 1.$$

Suppose that $\delta_n = (\log n)^\gamma/n$ for some $\gamma > 1$ and $\xi_{\min} = \Omega(\sqrt{\log n/(\log n)^\gamma})$, the conditions of theorem 5 are satisfied, and we have that the sequence of boundary fractions $\{\tau_j\}$ is consistently estimated. Since the boundary fractions are consistently estimated, we will see below that the estimated neighborhood $S(\hat{\boldsymbol{\theta}}^j)$ on the block $\hat{\mathcal{B}}^j$ consistently recovers the true neighborhood S^j.

Figure 1 illustrates the idea of correct neighborhood estimation on sufficiently large blocks.

Theorem 5. *Let $\{\mathbf{x}^i\}_{i\in[n]}$ be a sequence of observation according to the model in* (1). *Assume that the conditions of theorem 4 are satisfied. In addition, suppose that* **A5** *also holds. Then, if $|\hat{\mathcal{T}}| = B + 1$, it holds that*

$$\mathbb{P}[S^k = S(\hat{\boldsymbol{\theta}}^k)] \to 1, \qquad \forall k \in [B].$$

Under the assumptions of theorem 4 each estimated block is of size $\mathcal{O}(n)$. As a result, there are enough samples in each block to consistently estimate the underlying neighborhood structure. Observe that the neighborhood is consistently estimated at each $i \in \hat{\mathcal{B}}^j \cap \mathcal{B}^j$ for all $j \in [B]$ and the error is made only on the small fraction of samples, when $i \notin \hat{\mathcal{B}}^j \cap \mathcal{B}^j$, which is of order $\mathcal{O}(n\delta_n)$.

4 Discussion

We have addressed the problem of covariance selection when the underlying probability distribution changes abruptly at some points in time. Using a penalized neighborhood selection approach with the fused-type penalty, we are able to consistently estimate times when the distribution changes. Furthermore, our procedure estimates the network structure consistently whenever there is a large overlap between the estimated blocks and the unknown true blocks of samples coming from the same distribution. Applications of the proposed approach range from cognitive neuroscience, where the problem is to identify changing associations between different parts of a brain when presented with different stimuli, to system biology studies, where the task is to identify changing patterns of interactions between genes involved in different cellular processes.

References

[1] 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.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183202, 2009.

[3] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[4] 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.

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

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

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

[8] Mladen Kolar and Eric P Xing. Sparsistent estimation of Time-Varying discrete markov random fields. *0907.2337*, July 2009.

[9] Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.

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

[11] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, May 2005.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] 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.

[19] 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.

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

[21] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[22] 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.