Electronic Journal of Statistics
Vol. 6 (2012) 2069–2106
ISSN: 1935-7524
DOI: 10.1214/12-EJS739

Estimating networks with jumps

Mladen Kolar and Eric P. Xing*

Machine Learning Department
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
e-mail: mladenk@cs.cmu.edu; epxing@cs.cmu.edu

Abstract: We study the problem of estimating a temporally varying coefficient and varying structure (VCVS) graphical model underlying data collected over a period of time, such as social states of interacting individuals or microarray expression profiles of gene networks, as opposed to *i.i.d.* data from an invariant model widely considered in current literature of structural estimation. In particular, we consider the scenario in which the model evolves in a piece-wise constant fashion. We propose a procedure that estimates the structure of a graphical model by minimizing the temporally smoothed L1 penalized regression, which allows jointly estimating the partition boundaries of the VCVS model and the coefficient of the sparse precision matrix on each block of the partition. A highly scalable proximal gradient method is proposed to solve the resultant convex optimization problem; and the conditions for sparsistent estimation and the convergence rate of both the partition boundaries and the network structure are established for the first time for such estimators.

AMS 2000 subject classifications: Primary 62G05; secondary 62G20.
Keywords and phrases: Gaussian graphical models, network models, dynamic network models, structural changes.

Received February 2011.

Contents

1 Introduction 2070
2 Graph estimation via Temporal-Difference Lasso 2076
2.1 Numerical procedure 2078
2.2 Tuning parameter selection 2080
3 Theoretical results 2080
3.1 Assumptions 2081
3.2 Convergence of the partition boundaries 2082
3.3 Correct neighborhood selection 2084
4 Alternative estimation procedures 2085
4.1 Neighborhood selection with modified penalty 2085
4.2 Penalized maximum likelihood estimation 2086
5 Numerical studies 2087
6 Conclusion 2090

*This work is partially supported through the grants NIH R01GM087694 and AFOSR FA9550010247.

7 Proofs 2091
7.1 Proof of Lemma 1 2091
7.2 Proof of Theorem 2 2092
7.3 Proof of Lemma 4 2095
7.4 Proof of Theorem 5 2096
7.5 Proof of Lemma 6 2101
Acknowledgments 2101
Appendix 2101
References 2104

1. Introduction

Networks are a fundamental form of representation of relational information underlying large, noisy data from various domains. For example, in a biological study, nodes of a network can represent genes in one organism and edges can represent associations or regulatory dependencies among genes. In a social analysis, nodes of a network can represent actors and edges can represent interactions or friendships between actors. Exploring the statistical properties and hidden characteristics of network entities, and the stochastic processes behind temporal evolution of network topologies is essential for computational knowledge discovery and prediction based on network data.

In many dynamical environments, such as a developing biological system, it is often technically impossible to experimentally determine the network topologies specific to every time point in a certain time period. Resorting to computational inference methods, such as extant structural learning algorithms, is also difficult because for every model unique to a single time point, there exist as few as only a single snapshot of the nodal states distributed accordingly to the model in question. In this paper, we consider an estimation problem under a particular dynamic context, where the model evolves piecewise constantly, i.e., staying structurally invariant during unknown segments of time, and then jump to a different structure.

Approximately piecewise constantly evolving networks can be found underlying many natural dynamic systems of intellectual and practical interest. For example, in a biological developmental system such as the fruit fly, the entire life cycle of the fly consists of 4 discrete developmental stages, namely, embryo, larva, pupa, and adult; across the stages, one expect to see dramatical rewiring of the regulatory network to realize very different regulation functions due to different developmental needs, whereas within each stage, the change of the network topology are expected to be relatively more mild as revealed by the smoother trajectories of the gene expression activities, because a largely stable regulatory machinery is employed to control stage-specific developmental processes. Such phenomena are not uncommon in social systems. For example, in an underlying social network between the senators, even it is not visible to outsiders, we would imagine the network structure being more stable between the elections but more volatile when the campaigns start. Although it is legitimate

to use a completely unconstrained time-evolving network model to describe or analysis such systems, an approximately piecewise constantly evolving network model is better at capturing the different amount of network dynamics during different phases of a entire life cycle, and detecting boundaries between different phases when desirable.

A popular technique for deriving the network structure from *iid* sample is to estimate a sparse precision matrix. The importance of estimating precision matrices with zeros was recognized by Depmster [11] who coined the term *covariance selection*. The elements of the precision matrix represent the associations or conditional covariances between corresponding variables. Once a sparse precision matrix is estimated, a network can be drawn by connecting variables whose corresponding elements of the precision matrix are non-zero. Recent studies have shown that covariance selection methods based on the penalized likelihood maximization can lead to a consistent estimate of the network structure underlying a Gaussian Markov Random Fields [12, 32]. Moreover, a particular procedure for covariance selection known as neighborhood selection, which is built on ℓ_1 norm regularized regression, can produce a consistent estimate of the network structure when the sample is assumed to follow a general Markov Random Field distribution whose structure corresponds to the network in question [33, 28, 31]. Specifically, a Markov Random Field (MRF) is a probabilistic graphical model defined on a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is a vertex set corresponding to the set of random variables to be modeled (in this paper we call them *nodes* and *variables* interchangeably), and $E \subseteq V \times V$ is the edge set capturing conditional independencies among these nodes. Let $\mathbf{X} = (X_1, \ldots, X_p)'$ denote a p-dimensional random vector, whose elements are indexed by the nodes of the graph G. Under the MRF, a pair (a, b) is not an element of the edge set E if and only if the variable X_a is conditionally independent of X_b given all the rest of variables $X_{V\setminus\{a,b\}}$, $X_a \perp X_b | X_{V\setminus\{a,b\}}$. A distribution over $\mathbf{X}$ can be defined by taking the following log linear form that makes explicit use of the (presence and absence of edges in the) edge set: $p(\mathbf{X}) \propto \exp\{\sum_{(a,b)\in V} \theta_{ab} X_a X_b\}$. When the elements of the random vector $\mathbf{X}$ are discrete, e.g., $\mathbf{X} \in \{0, 1\}^p$, the model is referred to as a discrete MRF, sometimes known as an Ising model in statistics physics community; whereas when $\mathbf{X}$ is a continuous vector, the model is referred to as a Gaussian graphical model (GGM) because one can easily show that the $p(\mathbf{X})$ above is actually a multivariate Gaussian. The MRF have been widely used for modeling data with graphical relational structures over a fixed set of entities [39, 14]. The vertices can describe entities such as genes in a biological regulatory network, stocks in the market, or people in society; while the edges can describe relationships between vertices, for example, interaction, correlation or influence.

The statistical problem we concern in this paper is to estimate the structure of the Gaussian graphical model from observed samples of nodal states in a dynamic world. Traditional methods handle this problem with the assumption that the samples are *iid*. 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}), $1 \leq a, b \leq p$. Then one can obtain an estimator of the $\mathbf{\Omega}$ from $\mathcal{D}$ via optimizing a proper statistical loss function, such as likelihood or penalized likelihood. As mentioned earlier, 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. There has been a dramatic growth of interest in recent literature in the problem of covariance selection, which deals with the graph estimation problem above. Existing works range from algorithmic development focusing on efficient estimation procedures, to theoretical analysis focusing on statistical guarantees of different estimators. We do not intend to give an extensive overview of the literature here, but interested readers can follow the pointers bellow. In the classical literature (e.g., [22]), procedures are developed for small dimensional graphs and commonly involve hypothesis testing with greedy selection of edges. More recent literature estimates the sparse precision matrix by optimizing penalized likelihood [42, 12, 4, 35, 13, 32, 16, 44] or through neighborhood selection [28, 31, 15, 40], 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 a single invariant network model is sufficient to describe the dependencies in the observed data. However, when the observed data are not *iid*, such an assumption is not justifiable. For example, when data consist of microarray measurements of the gene expression levels collected throughout the cell cycle or development of an organism, different genes can be active during different stages. This suggests that different distributions and hence different networks should be used to describe the 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. This semi-parametric generalization of the parametric model is referred to as a varying coefficient varying structure (VCVS) model.

Now, let $\{\mathbf{x}_i\}_{i\in[n]} \in \mathbb{R}^p$ be a sequence of n *conditionally* independent observations [1] (we use $[n]$ to denote the set $\{1, \ldots, n\}$) from some p-dimensional multivariate normal distributions, 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 $\mathcal{T} := \{T_0 = 1 < T_1 < \cdots < T_B = n + 1\}$ denote the set of partition boundaries.

[1] We emphasize that the independence is only present when *each* instance of the latent time varying model is given. In practice, such models are unknown, and therefore marginally the samples are dependent. Furthermore, the instances of the latent evolving models generating the samples are NOT independent, as we can see in later presentation.

We consider the following model

$$\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}^j), \qquad i \in \mathcal{B}^j, \tag{1.1}$$

such 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 $\mathbf{\Sigma}^j = (\sigma^j_{ab})_{a,b\in[p]}$, which suggest that it is only unique to segment j of the time series. 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 both the partition set $\{\mathcal{B}^j\}$ and the non-zero elements of the precision matrices $\{\mathbf{\Omega}^j\}_{j\in[B]}$ from the sample $\{\mathbf{x}_i\}_{i\in[n]}$. Note that in this work we study a particular case of the VCVS model, where the coefficients are piece-wise constant functions of time. Although this model does not yet entirely agree with how a real world time series data would behave, as none existing model does, this instantiation of the VCVS model come one step closer in some sense to the real world scenario than other popular approaches for time series modeling, such as Hidden Markov Models or state space models, where stationary emission models such as linear Gaussian are usually employed to relate observation at different time points to simple latent states. Here, instead of positing an observation at time t to be derived from a latent state that transitions stationarily from a previous time point, we assume that such an observation is generated from a latent network model that are related to the network models active at the previous and subsequent time points nonparametrically. As suggested in the introduction, many real world dynamic systems, such as the stage-specific development of multi-cellular organisms like the fruit fly, and the evolving network of latent relatedness between politicians, are likely to behave approximately piecewise constantly; therefore time series data from such systems, such as the continuous-valued gene expression microarray time series, and the discrete-valued voting records, are suitable examples where our proposed models can be applied to [1]. A scenario where the coefficients are smoothly varying functions of time has been considered in [44] for the GGM and in [21] and [19] for an Ising model, which complements the model studied in this paper, whose asymptotic properties are somewhat easier to analyze as we have shown earlier.

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 model estimates are forced to stay similar across time using a fusion-type total variation penalty and the sparsity of each neighborhood is obtained through the ℓ_1 penalty. Details of the approach are given in §2.

The model in Eq. (1.1) is related to the varying-coefficient models (e.g., [18]) with the coefficients being piece-wise constant functions. Varying coefficient regression models with piece-wise constant coefficients are also known as segmented multivariate regression models [24] or linear models with structural changes [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.

Little work has been done so far towards modeling dynamic networks and estimating changing precision matrices. [44] develops a nonparametric method for estimation of time-varying GGM, where $\mathbf{x}^t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}(t))$ and $\mathbf{\Sigma}(t)$ is smoothly changing over time. The procedure is based on the penalized likelihood approach of [42] with the empirical covariance matrix obtained using a kernel smoother. Our work is very different from the one of [44], since under our assumptions the network changes abruptly rather than smoothly. Furthermore, as we outline in §2, our estimation procedure is not based on the penalized likelihood approach. Estimation of time-varying Ising models has been discussed in [1] and [21]. [41] and [20] studied nonparametric ways to estimate the conditional covariance matrix. The work of [1] is most similar to our setting, where they also use a fused-type penalty combined with an ℓ_1 penalty to estimate the structure of the varying Ising model. Here, in addition to focusing on GGMs, there is an additional subtle, but important, difference to [1]. In this work, we use a modification of the fusion penalty (formally described in §2) which allows us to characterize the model selection consistency of our estimates and the convergence properties of the estimated partition boundaries, which is not available in the earlier work.

The remaining of the paper is organized as follows. In §2, we describe our estimation procedure and provide an efficient first-order optimization procedure capable of estimating large graphs. The optimization procedure is based on the smoothing procedure of [29] and converges in $\mathcal{O}(1/\epsilon)$ iterations, where ϵ is the desired accuracy. Our main theoretical results are presented in §3. In particular, we show that the partition boundaries are estimated consistently. Furthermore, the graph structure is consistently estimated on every block of the partition that contains enough samples. In §4, we discuss alternative estimation procedures based on penalized maximum likelihood estimation, instead of the neighborhood selection. Numerical studies showing the finite sample performance of our procedure are given in §5. The proofs of the main results are relegated to §7, with some technical details presented in Appendix.

Notation schemes

For clarity, we end the introduction with a summary of the notations used in the paper. We use $[n]$ to denote the set $\{1, \ldots, n\}$ and $[l : r]$ to denote the set $\{l, l+1, \ldots, r-1\}$. We use $\mathcal{B}^j$ to denote j-th block of the partition $\mathcal{T}$. With some abuse of notation, we also use $\mathcal{B}^j$ to denote the set $[T_{j-1} : T_j]$. The number of samples in the block $\mathcal{B}^j$ is denoted as $|\mathcal{B}^j|$. For a set $S \subset V$, we use the notation X_S to denote the set $\{X_a : a \in S\}$ of random variables. We use $\mathbf{X}$ to denote the $n \times p$ matrix whose rows consist of observations. The vector $\mathbf{X}_a = (x_{1,a}, \ldots, x_{n,a})'$ denotes a column of matrix $\mathbf{X}$ and, similarly, $\mathbf{X}_S = (\mathbf{X}_b : b \in S)$ denotes the $n \times |S|$ sub-matrix of $\mathbf{X}$ whose columns are indexed by the set S and $\mathbf{X}^{\mathcal{B}^j}$ denotes the sub-matrix $|\mathcal{B}^j| \times p$ whose rows are

TABLE 1
Summary of symbols used throughout the paper

Symbol	Meaning	Example
$[n]$	used to denote the set $\{1,\dots,n\}$	
$[t_1:t_2]$	used to denote the set $\{t_1,t_1+1,\dots,t_2-1\}$	
i	used for indexing related to samples	$\mathbf{x}^i$ or $\beta^a_{\cdot,i}$
j,k	used for indexing related to block	$\boldsymbol{\theta}^{a,j}$ or S^k_a
a,b	used for indexing nodes in a graph	$a,b\in V$
G	the graph consisting of vertices and edges	$G=(V,E)$
V	the set of nodes in a graph	$V=[p]$
E_i	the set of edges at time i	
X_a	the component of a random vector $\mathbf{X}$ indexed by the vertex a	
$\boldsymbol{\beta}^a_{\cdot,i}$	the vector of regression coefficients for sample i	
$\boldsymbol{\theta}^{a,j}$	the vector of regression coefficients for block j	
$\mathcal{T}$	the set of partition boundaries	
$\{\tau_j\}_j$	the set of boundary fractions	$T_j=\lfloor n\tau_j\rfloor$
$\mathcal{B}^j$	an index set for the samples in the partition j	$\mathcal{B}^j\subset[n]$
B	denotes the number of partitions	
S^j_a	the set of neighbors of node a in block j	
$S(\boldsymbol{\theta}^{a,j})$	the set of non-zero elements of $\boldsymbol{\theta}^{a,j}$	
$\bar{S}^j_a$	the closure of S^j_a	$\bar{S}^j_a=S^j_a\cup\{a\}$
N^j_a	nodes not in the neighborhood of the node a in block j	$N^j_a=[p]\backslash\bar{S}^j_a$
$\backslash a$	the set of all vertices excluding the vertex a	$\backslash a=[p]\backslash\{a\}$
$\lvert\cdot\rvert$	cardinality of a set or absolute value	
$\boldsymbol{\Sigma}$	the covariance matrix	
σ_{ab}	an element of the covariance matrix	
$\boldsymbol{\Omega}$	the precision matrix	
ω_{ab}	an element of the precision matrix	
$\langle\cdot,\cdot\rangle$	the dot product	$\langle\mathbf{a},\mathbf{b}\rangle=\mathbf{a}'\mathbf{b}$
$\langle\!\langle\cdot,\cdot\rangle\!\rangle$	the dot product between matrices	$\langle\!\langle\mathbf{A},\mathbf{B}\rangle\!\rangle=\mathrm{tr}(\mathbf{A}'\mathbf{B})$
$\xi_{\min}$	the minimum change between regression coefficient	$\lVert\boldsymbol{\theta}^{a,j}-\boldsymbol{\theta}^{a,j-1}\rVert_2\geq\xi_{\min}$
$\theta_{\min}$	the minimum size of a coefficient	$\lvert\theta^{a,j}_b\rvert\geq\theta_{\min}$
$\Delta_{\min}$	the minimum size of a block	$\lvert\mathcal{B}^j\rvert\geq\Delta_{\min}$

indexed by the set $\mathcal{B}^j$. For simplicity of notation, we will use $\backslash a$ to denote the index set $[p]\backslash\{a\}$, $\mathbf{X}_{\backslash a}=(\mathbf{X}_b\ :\ b\in[p]\backslash\{a\})$. For a vector $\mathbf{a}\in\mathbb{R}^p$, we let $S(\mathbf{a})$ denote the set of non-zero components of $\mathbf{a}$. Throughout the paper, we use $c_1,c_2,\dots$ to denote positive constants whose value may change from line to line. For a vector $\mathbf{a}\in\mathbb{R}^n$, define $||\mathbf{a}||_1=\sum_{i\in[n]}|a_i|$, $||\mathbf{a}||_2=\sqrt{\sum_{i\in[n]}a_i^2}$ and $||\mathbf{a}||_\infty=\max_i|a_i|$. For a symmetric matrix $\mathbf{A}$, $\Lambda_{\min}(\mathbf{A})$ denotes the smallest and $\Lambda_{\max}(\mathbf{A})$ the largest eigenvalue. For a matrix $\mathbf{A}$ (not necessarily symmetric), we use $|\!|\!|\mathbf{A}|\!|\!|_\infty=\max_i\sum_j|A_{ij}|$. For two vectors $\mathbf{a},\mathbf{b}\in\mathbb{R}^n$, the dot product is denoted $\langle\mathbf{a},\mathbf{b}\rangle=\sum_{i\in[n]}a_ib_i$. For two matrices $\mathbf{A},\mathbf{B}\in\mathbb{R}^{n\times m}$, the dot product is denoted as $\langle\!\langle\mathbf{A},\mathbf{B}\rangle\!\rangle=\mathrm{tr}(\mathbf{A}'\mathbf{B})$. Given two sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n=\mathcal{O}(b_n)$ means that there exists a constant c_1 such that $a_n\leq c_1b_n$; the notation $a_n=\Omega(b_n)$ means that there exists a constant c_2 such that $a_n\geq c_2b_n$ and the notation $a_n\asymp b_n$ means that $a_n=\mathcal{O}(b_n)$ and $b_n=\mathcal{O}(a_n)$. Similarly, we will use the notation $a_n=o_p(b_n)$ to denote that $b_n^{-1}a_n$ converges to 0 in probability.

2. Graph estimation via Temporal-Difference Lasso

In this section, we introduce our time-varying covariance selection procedure, which is based on the time-coupled neighborhood selection using the fused-type penalty. We call the proposed procedure Temporal-Difference Lasso (*TD-Lasso*). We start by reviewing the basic neighborhood selection procedure, which has previously been used to estimate graphs in, e.g., [31, 28, 33, 15].

We start by relating 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 $\mathbf{\Omega}$, $S_a = \{b \in [p] \backslash \{a\} \ : \ \omega_{ab} \neq 0\}$. See [22] for more details. It is a well known result for Gaussian graphical models that the elements of

$$\boldsymbol{\theta}^a = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \ \mathbb{E}\left(X_a - \sum_{b \in \backslash a} X_b \theta_b \right)^2$$

are given by $\theta_b^a = -\omega_{ab}/\omega_{aa}$. Therefore, the neighborhood of a node a, S_a, is equal to the set of non-zero coefficients of $\boldsymbol{\theta}^a$. Using the expression for $\boldsymbol{\theta}^a$, we can write $X_a = \sum_{b \in S_a} X_b \theta_b^a + \epsilon$, where ϵ is independent of $X_{\backslash a}$.

The neighborhood selection procedure was motivated by the above relationship between the regression coefficients and the elements of the precision matrix. [28] proposed to solve the following optimization procedure

$$\hat{\boldsymbol{\theta}}^a = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \ \frac{1}{n} ||\mathbf{X}_a - \mathbf{X}_{\backslash a} \boldsymbol{\theta}||_2^2 + \lambda ||\boldsymbol{\theta}||_1 \tag{2.1}$$

and proved that for *iid* sample the non-zero coefficients of $\hat{\boldsymbol{\theta}}^a$ consistently estimate the neighborhood of the node a, under a suitably chosen penalty parameter λ.

In this paper, we build on the neighbourhood selection procedure to estimate the changing graph structure in model (1.1). We use S_a^j 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, $N_a^j = V \backslash S_a^j$. Consider the following estimation procedure

$$\hat{\boldsymbol{\beta}}^a = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p-1 \times n}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{\beta}) + \text{pen}_{\lambda_1, \lambda_2}(\boldsymbol{\beta}) \tag{2.2}$$

where the loss is defined for $\boldsymbol{\beta} = (\beta_{b,i})_{b \in [p-1], i \in [n]}$ as

$$\mathcal{L}(\boldsymbol{\beta}) := \sum_{i \in [n]} \left(x_{i,a} - \sum_{b \in \backslash a} x_{i,b} \beta_{b,i} \right)^2 \tag{2.3}$$

and the penalty is defined as

$$\text{pen}_{\lambda_1, \lambda_2}(\boldsymbol{\beta}) := 2\lambda_1 \sum_{i=2}^{n} ||\boldsymbol{\beta}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i-1}||_2 + 2\lambda_2 \sum_{i=1}^{n} \sum_{b \in \backslash a} |\beta_{b,i}|. \tag{2.4}$$

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 within 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

$$\hat{S}^j_a := S(\hat{\boldsymbol{\theta}}^{a,j}) := \{b \in \backslash a \ : \ \hat{\theta}^{a,j}_b \neq 0\}.$$

Solving (2.2) 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 [34, 27], 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 [36], 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.

The objective (2.2) estimates the neighborhood of one node in a graph for all time-points. After solving the objective (2.2) for all nodes $a \in V$, we need to combine them to obtain the graph structure. We will use the following procedure to combine $\{\hat{\boldsymbol{\beta}}^a\}_{a \in V}$,

$$\hat{E}_i = \{(a,b) \ : \ \max(|\beta^a_{b,i}|, |\beta^b_{a,i}|) > 0\}, \qquad i \in [n].$$

That is, an edge between nodes a and b is included in the graph if at least one of the nodes a or b is included in the neighborhood of the other node. We use the max operator to combine different neighborhoods as we believe that for the purpose of network exploration it is more important to occasionally include spurious edges than to omit relevant ones. For further discussion on the differences between the min and the max combination, we refer an interested reader to [4].

2.1. Numerical procedure

Finding a minimizer $\hat{\boldsymbol{\beta}}^a$ of (2.2) can be a computationally challenging task for an off-the-shelf convex optimization procedure. We propose to use an accelerated gradient method with a smoothing technique [29], 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

$$H_{ij} = \begin{cases} -1 & \text{if } i = j \\ 1 & \text{if } i = j+1 \\ 0 & \text{otherwise.} \end{cases}$$

With the matrix $\mathbf{H}$ we can rewrite the fused penalty term as $2\lambda_1 \sum_{i=1}^{n-1} ||(\boldsymbol{\beta}\mathbf{H})_{\cdot,i}||_2$ and using the fact that the ℓ_2 norm is self dual (e.g., see [7]) we have the following representation

$$2\lambda_1 \sum_{i=2}^{n} ||\boldsymbol{\beta}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i-1}||_2 = \max_{\mathbf{U} \in \mathcal{Q}} \langle\langle \mathbf{U}, 2\lambda_1 \boldsymbol{\beta}\mathbf{H} \rangle\rangle \tag{2.5}$$

where $\mathcal{Q} := \{\mathbf{U} \in \mathbb{R}^{p-1 \times n-1} \ : \ ||\mathbf{U}_{\cdot,i}||_2 \leq 1, \ \forall i \in [n-1]\}$. The following function is defined as a smooth approximation to the fused penalty,

$$\Psi_\mu(\boldsymbol{\beta}) := \max_{\mathbf{U} \in \mathcal{Q}} \langle\langle \mathbf{U}, 2\lambda_1 \boldsymbol{\beta}\mathbf{H} \rangle\rangle - \mu ||\mathbf{U}||_F^2 \tag{2.6}$$

where $\mu > 0$ is the smoothness parameter. It is easy to see 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. Let $\mathbf{U}_\mu(\boldsymbol{\beta})$ be the optimal solution of the maximization problem in (2.6), which can be obtained analytically as

$$\mathbf{U}_\mu(\boldsymbol{\beta}) = \Pi_{\mathcal{Q}} \left(\frac{\lambda \boldsymbol{\beta}\mathbf{H}}{\mu} \right) \tag{2.7}$$

where $\Pi_{\mathcal{Q}}(\cdot)$ is the projection operator onto the set $\mathcal{Q}$. From Theorem 1 in [29], we have that $\Psi_\mu(\boldsymbol{\beta})$ is continuously differentiable and convex, with the gradient

$$\nabla \Psi_\mu(\boldsymbol{\beta}) = 2\lambda_1 \mathbf{U}_\mu(\boldsymbol{\beta})\mathbf{H}' \tag{2.8}$$

that is Lipschitz continuous.

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 [5] (see also [30]), we define the following quadratic approximation of $F(\boldsymbol{\beta})$ at a point $\boldsymbol{\beta}_0$

$$\begin{aligned} Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}_0) := \mathcal{L}(\boldsymbol{\beta}_0) + \Psi_\mu(\boldsymbol{\beta}_0) + \langle\langle \boldsymbol{\beta} - \boldsymbol{\beta}_0, \nabla\mathcal{L}(\boldsymbol{\beta}_0) + \nabla\Psi(\boldsymbol{\beta}_0)\rangle\rangle \\ + \frac{L}{2}||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_F^2 + 2\lambda_2||\boldsymbol{\beta}||_1 \end{aligned} \tag{2.9}$$

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 a minimizer of $Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$. Ignoring constant terms, $p_L(\boldsymbol{\beta}_0)$ can be obtained as

$$p_L(\boldsymbol{\beta}_0) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1\times n}}{\operatorname{argmin}} \ \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{\beta}_0 - \frac{1}{L}\big(\nabla\mathcal{L} + \nabla\Psi\big)(\boldsymbol{\beta}_0)\right)\right\|_F^2 + \frac{2\lambda_2}{L}\|\boldsymbol{\beta}\|_1.$$

It is clear that $p_L(\boldsymbol{\beta}_0)$ is the unique minimizer, which can be obtained in a closed form, as a result of the soft-thresholding,

$$p_L(\boldsymbol{\beta}_0) = T\left(\boldsymbol{\beta}_0 - \frac{1}{L}\big(\nabla\mathcal{L} + \nabla\Psi\big)(\boldsymbol{\beta}_0), \frac{2\lambda_2}{L}\right) \tag{2.10}$$

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

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. Combining all of the above, we arrive at Algorithm 1. In the algorithm, $\boldsymbol{\beta}_0$ is set to zero or, if the optimization problem is solved for a sequence of tuning parameters, it can be set to the solution $\hat{\boldsymbol{\beta}}$ obtained for the previous set of tuning parameters. The parameter γ is a constant used to increase the estimate

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. (2.10))
  $\alpha_{k+1} := \frac{1+\sqrt{1+4\alpha_k}}{2}$
  $\mathbf{z}_{k+1} := \boldsymbol{\beta}_k + \frac{\alpha_k - 1}{\alpha_{k+1}}\big(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\big)$
until *convergence*
$\hat{\boldsymbol{\beta}}^a := \boldsymbol{\beta}_k$

Algorithm 1: Accelerated Gradient Method for Equation (2.2)

of the Lipschitz constant L and we set it to $\gamma = 1.5$ in our experiments, while $L = 1$ initially. Compared to the gradient descent method (which can be obtain by iterating $\boldsymbol{\beta}_{k+1} = p_L(\boldsymbol{\beta}_k)$), the accelerated gradient method updates two sequences $\{\boldsymbol{\beta}_k\}$ and $\{\mathbf{z}_k\}$ recursively. Instead of performing the gradient step from the latest approximate solution $\boldsymbol{\beta}_k$, the gradient step is performed from the search point $\mathbf{z}_k$ that is obtained as a linear combination of the last two approximate solutions $\boldsymbol{\beta}_{k-1}$ and $\boldsymbol{\beta}_k$. Since the condition $F(p_L(\mathbf{z}_k)) \leq Q_L(p_L(\mathbf{z}_k), \mathbf{z}_k)$ is satisfied in every iteration, we have the algorithm converges in $\mathcal{O}(1/\epsilon)$ iterations following [5]. As the convergence criterion, we stop iterating once the relative change in the objective value is below some threshold value (e.g., we use 10^{-4}).

2.2. Tuning parameter selection

The penalty parameters λ_1 and λ_2 control the complexity of the estimated model. In this work, we propose to use the BIC score to select the tuning parameters. Define the BIC score for each node $a \in V$ as

$$\mathrm{BIC}_a(\lambda_1, \lambda_2) := \log \frac{\mathcal{L}(\hat{\boldsymbol{\beta}}^a)}{n} + \frac{\log n}{n} \sum_{j \in [\hat{B}]} |S(\hat{\boldsymbol{\theta}}^{a,j})| \tag{2.11}$$

where $\mathcal{L}(\cdot)$ is defined in (2.3) and $\hat{\boldsymbol{\beta}}^a = \hat{\boldsymbol{\beta}}^a(\lambda_1, \lambda_2)$ is a solution of (2.2). The penalty parameters can now be chosen as

$$\{\hat{\lambda}_1, \hat{\lambda}_2\} = \operatorname*{argmin}_{\lambda_1, \lambda_2} \sum_{a \in V} \mathrm{BIC}_a(\lambda_1, \lambda_2). \tag{2.12}$$

We will use the above formula to select the tuning parameters in our simulations, where we are going to search for the best choice of parameters over a grid.

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 an asymptotic framework by letting the sample size n grow, while keeping the other parameters fixed. 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, if the number of partition blocks is estimated correctly, then we show that $\max_{j \in [B]} |\hat{T}_j - T_j| \leq n\delta_n$ with probability tending to 1, where $\{\delta_n\}_n$ is a non-increasing sequence of positive numbers that tends to zero. If the number of partition segments is over estimated, then we show that for a distance defined for two sets A and B as

$$h(A, B) := \sup_{b \in B} \inf_{a \in A} |a - b|, \tag{3.1}$$

we have $h(\hat{\mathcal{T}}, \mathcal{T}) \leq n\delta_n$ with probability tending to 1. 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.

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 (2.2) need to satisfy. The optimality conditions for $\hat{\boldsymbol{\beta}}^a$ to be a solution of (2.2) 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 satisfies $h(\hat{\mathcal{T}}, \mathcal{T}) \geq n\delta_n$. 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 (2.2) result in partition boundaries that are "close" to the true partition boundaries, with high-probability. Once it is established that $\hat{\mathcal{T}}$ and $\mathcal{T}$ satisfy $h(\hat{\mathcal{T}}, \mathcal{T}) \leq n\delta_n$, we can further show that the neighborhood estimates are consistently estimated, under the assumption that the estimated blocks of the partition have enough samples. This part of the analysis follows the commonly used strategy to prove that the Lasso is sparsistent (e.g., see [9, 38, 28]), however important modifications are required due to the fact that position of the partition boundaries are being estimated.

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, $\xi_{\min} := \min_{a\in V} \min_{j\in[B-1]} \times ||\boldsymbol{\theta}^{a,j+1} - \boldsymbol{\theta}^{a,j}||_2$ denote the minimum jump size and $\theta_{\min} = \min_{a\in V} \min_{j\in[B]} \times \min_{b\in S^j} |\theta_b^{a,j}|$ 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

$$\phi_{\min} = \min \{\Lambda_{\min}(\boldsymbol{\Sigma}^j) \ : \ j \in [B], a \in V\}$$

and

$$\phi_{\max} = \max \{\Lambda_{\max}(\boldsymbol{\Sigma}^j) \ : \ j \in [B], a \in V\}.$$

A2 Variables are scaled so that $\sigma_{aa}^j = 1$ 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 $\alpha \in (0, 1]$, such that the following holds

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

The assumption **A4** states that the variables in the neighborhood of the node a, S_a^j, are not too correlated with the variables in the set N_a^j. This assumption is necessary and sufficient for correct identification of the relevant variables in the Lasso regression problems (e.g., see [43, 37]). Note that this condition is sufficient also in our case when the correct partition boundaries are not known.

A5 The minimum coefficient size $\theta_{\min}$ satisfies $\theta_{\min} = \Omega(\sqrt{\log(n)/n})$.

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.

A6 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. Convergence of the partition boundaries

In this subsection we establish the rate of convergence of the boundary partitions for the estimator (2.2). We start by giving a lemma that characterizes solutions of the optimization problem given in (2.2). Note that the optimization problem in (2.2) is convex, however, there may be multiple solutions to it, since it is not strictly convex.

Lemma 1. *Let* $x_{i,a} = \mathbf{x}'_{i,\setminus a}\boldsymbol{\theta}_a + \epsilon_i$. *A matrix* $\hat{\boldsymbol{\beta}}$ *is optimal for the optimization problem* (2.2) *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*

$$\sum_{i=k}^{n} \mathbf{x}_{i,\setminus a} \langle \mathbf{x}_{i,\setminus a}, \hat{\boldsymbol{\beta}}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i} \rangle - \sum_{i=k}^{n} \mathbf{x}_{i,\setminus a} \epsilon_i + \lambda_1 \hat{\mathbf{z}}_k + \lambda_2 \sum_{i=k}^{n} \hat{\mathbf{y}}_i = 0 \tag{3.2}$$

for all $k \in [n]$ *and* $\hat{\mathbf{z}}_1 = \hat{\mathbf{z}}_{n+1} = \mathbf{0}$.

The following theorem provides the convergence rate of the estimated boundaries of $\hat{\mathcal{T}}$, under the assumption that the correct number of blocks is known.

Theorem 2. *Let $\{\mathbf{x}_i\}_{i\in[n]}$ be a sequence of observation according to the model in (1.1). Assume that* **A1**-**A3** *and* **A5**-**A6** *hold. Suppose that the penalty parameters λ_1 and λ_2 satisfy*

$$\lambda_1 \asymp \lambda_2 = \mathcal{O}(\sqrt{\log(n)/n}). \tag{3.3}$$

Let $\{\hat{\boldsymbol{\beta}}_{\cdot,i}\}_{i\in[n]}$ be any solution of (2.2) and let $\hat{\mathcal{T}}$ be the associated estimate of the block partition. 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 $(n\delta_n\xi_{\min})^{-1}\lambda_1 \to 0$, $\xi_{\min}^{-1}\sqrt{p}\lambda_2 \to 0$ and $(\xi_{\min}\sqrt{n\delta_n})^{-1}\sqrt{p\log n} \to 0$, then if $|\hat{\mathcal{T}}| = B+1$ the following holds

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

The proof builds on techniques developed in [17] and is presented in §7.

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 2 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.

Unfortunately, the correct bound on the number of block B may not be known. However, a conservative upper bound $B_{\max}$ on the number of blocks B may be known. Suppose that the sequence of observation is over segmented, with the number of estimated blocks bounded by $B_{\max}$. Then the following proposition gives an upper bound on $h(\hat{\mathcal{T}}, \mathcal{T})$ where $h(\cdot,\cdot)$ is defined in (3.1).

Proposition 3. *Let $\{\mathbf{x}_i\}_{i\in[n]}$ be a sequence of observation according to the model in (1.1). Assume that the conditions of Theorem 2 are satisfied. Let $\hat{\boldsymbol{\beta}}$ be a solution of (2.2) and $\hat{\mathcal{T}}$ the corresponding set of partition boundaries, with $\hat{B}$ blocks. If the number of blocks satisfy $B \leq \hat{B} \leq B_{\max}$, then*

$$\mathbb{P}[h(\hat{\mathcal{T}}, \mathcal{T}) \leq n\delta_n] \xrightarrow{n\to\infty} 1.$$

The proof of the proposition follows the same ideas of Theorem 2 and its sketch is given in the appendix.

The above proposition assures us that even if the number of blocks is overestimated, there will be a partition boundary close to every true unknown partition boundary. In many cases it is reasonable to assume that a practitioner would have an idea about the number of blocks that she wishes to discover. In that way, our procedure can be used to explore and visualize the data. It is still an open question to pick the tuning parameters in a data dependent way so that the number of blocks are estimated consistently.

FIG 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.*

3.3. Correct neighborhood selection

In this section, we give a result on the consistency of the neighborhood estimation. We will show that whenever the estimated block $\hat{\mathcal{B}}^j$ is large enough, say $|\hat{\mathcal{B}}^j| \geq r_n$ where $\{r_n\}_{n\geq 1}$ is 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 $n \to \infty$, we have that $S(\hat{\boldsymbol{\theta}}^j) = S(\boldsymbol{\beta}^k)$, where $\boldsymbol{\beta}^k$ is the true parameter on the true block $\mathcal{B}^k$ that overlaps $\hat{\mathcal{B}}^j$ the most. Figure 1 illustrates this idea. The blue region in the figure denotes the overlap between the true block and the estimated block of the partition. The orange region corresponds to the overlap of the estimated block with a different true block. If the blue region is considerably larger than the orange region, the bias coming from the sample from the orange region will not be strong enough to disable us from selecting the correct neighborhood. On the other hand, since the orange region is small, as seen from Theorem 2, there is little hope of estimating the neighborhood correctly on that portion of the sample.

Suppose that we know that there is a solution to the optimization problem (2.2) with the partition boundary $\hat{\mathcal{T}}$. Then that solution is also a minimizer of the following objective

$$\min_{\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{\hat{B}}} \sum_{j\in\hat{B}} ||\mathbf{X}_a^{\hat{\mathcal{B}}^j} - \mathbf{X}_{\setminus a}^{\hat{\mathcal{B}}^j}\boldsymbol{\theta}^j||_2^2 + 2\lambda_1 \sum_{j=2}^{\hat{B}} ||\boldsymbol{\theta}^j - \boldsymbol{\theta}^{j-1}||_2 + 2\lambda_2 \sum_{j=1}^{\hat{B}} |\hat{\mathcal{B}}^j| ||\boldsymbol{\theta}^j||_1. \quad (3.4)$$

Note that the problem (3.4) does not give a practical way of solving (2.2), but will help us to reason about the solutions of (2.2). In particular, while there may be multiple solutions to the problem (2.2), under some conditions, we can characterize the sparsity pattern of any solution that has specified partition boundaries $\hat{\mathcal{T}}$.

Lemma 4. *Let $\hat{\boldsymbol{\beta}}$ be a solution to (2.2), 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 pattern of a solution with the partition boundary $\hat{\mathcal{T}}$ is unique. Note, however, that there may other solutions to (2.2) that have different partition boundaries.

Now, we are ready to state the following theorem, which establishes that the correct neighborhood is selected on every sufficiently large estimated block of the partition.

Theorem 5. *Let $\{\mathbf{x}_i\}_{i\in[n]}$ be a sequence of observation according to the model in (1.1). Assume that the conditions of theorem 2 are satisfied. In addition, suppose that* **A4** *also holds. Then, if $|\hat{\mathcal{T}}| = B + 1$, it holds that*

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

Under the assumptions of theorem 2 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 $\mathcal{O}(n\delta_n)$.

Using proposition 3 in place of theorem 2, it can be similarly shown that, for a large fraction of samples, the neighborhood is consistently estimated even in the case of over-segmentation. In particular, whenever there is a sufficiently large estimated block, with $|\hat{\mathcal{B}}^k \cap \mathcal{B}^j| = \mathcal{O}(r_n)$, it holds that $S(\hat{\mathcal{B}}^k) = S^j$ with probability tending to one.

4. Alternative estimation procedures

In this section, we discuss some alternative estimation methods to the neighborhood selection detailed in §2. We start describing how to solve the objective (2.2) for different penalties than the one given in (2.4). In particular, we describe how to minimize the objective when the ℓ_2 is replaced with the ℓ_q ($q \in \{1, \infty\}$) norm in (2.4). Next, we describe how to solve the penalized maximum likelihood objective with the temporal difference penalty. We do not provide statistical guarantees for solutions of these objective functions.

4.1. Neighborhood selection with modified penalty

We consider the optimization problem given in (2.2) with the following penalty

$$\text{pen}_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) := 2\lambda_1 \sum_{i=2}^{n} ||\boldsymbol{\beta}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i-1}||_q + 2\lambda_2 \sum_{i=1}^{n} \sum_{b\in\setminus a} |\beta_{b,i}|, \qquad q \in \{1, \infty\}. \tag{4.1}$$

We call the penalty in (4.1) the TD_q penalty. As in §2.1, we apply the smoothing procedure to the first term in (4.1). Using the dual norm representation, we have

$$2\lambda_1 \sum_{i=2}^{n} ||\boldsymbol{\beta}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i-1}||_q = \max_{\mathbf{U}\in\mathcal{Q}^q} \langle\langle \mathbf{U}, 2\lambda_1 \boldsymbol{\beta}\mathbf{H} \rangle\rangle$$

where

$$\mathcal{Q}^1 := \{\mathbf{U} \in \mathbb{R}^{p-1 \times n-1} \ : \ ||\mathbf{U}_{\cdot,i}||_\infty \leq 1, \ \forall i \in [n-1]\}$$

and

$$\mathcal{Q}^\infty := \{\mathbf{U} \in \mathbb{R}^{p-1 \times n-1} \ : \ ||\mathbf{U}_{\cdot,i}||_1 \leq 1, \ \forall i \in [n-1]\}.$$

Next, we define smooth approximation to the norm as

$$\Psi_\mu^q(\boldsymbol{\beta}) := \max_{\mathbf{U} \in \mathcal{Q}^q} \langle\langle \mathbf{U}, 2\lambda_1 \boldsymbol{\beta}\mathbf{H} \rangle\rangle - \mu ||\mathbf{U}||_F^2 \tag{4.2}$$

where $\mu > 0$ is the smoothness parameter. Let

$$\mathbf{U}_\mu^q(\boldsymbol{\beta}) = \Pi_{\mathcal{Q}^q}\left(\frac{\lambda \boldsymbol{\beta}\mathbf{H}}{\mu}\right) \tag{4.3}$$

be the optimal solution of the maximization problem in (4.2), where $\Pi_{\mathcal{Q}^q}(\cdot)$ is the projection operator onto the set $\mathcal{Q}^q$. We observe that the projection on the ℓ_∞ unit ball can be easily obtained, while a fast algorithm for projection on the ℓ_1 unit ball can be found in [8]. The gradient can now be obtained as

$$\nabla \Psi_\mu^q(\boldsymbol{\beta}) = 2\lambda_1 \mathbf{U}_\mu^q(\boldsymbol{\beta})\mathbf{H}', \tag{4.4}$$

and we can proceed as in § 2.1 to arrive at the update (2.10).

We have described how to optimize (2.2) with the TD_q penalty for $q \in \{1, 2, \infty\}$. Other ℓ_q norms are not commonly used in practice. We also note that a different procedure for $q = 1$ can be found in [26].

4.2. Penalized maximum likelihood estimation

In §2, we have related the problem of estimating zero elements of a precision matrix to a penalized regression procedure. Now, we consider estimating a sparse precision matrix using a penalized maximum likelihood approach. That is, we consider the following optimization procedure

$$\min_{\{\boldsymbol{\Omega}_i \succ \mathbf{0}\}_{i \in [n]}} \sum_{i \in [n]} (\text{tr}\, \boldsymbol{\Omega}_i \mathbf{x}_i \mathbf{x}_i' - \log |\boldsymbol{\Omega}_i|) + \text{pen}_{\lambda_1, \lambda_2}(\{\boldsymbol{\Omega}_t\}_{t \in [n]}) \tag{4.5}$$

where

$$\text{pen}_{\lambda_1, \lambda_2}(\{\boldsymbol{\Omega}_i\}_{i \in [n]}) := 2\lambda_1 \sum_{i=2}^n ||\boldsymbol{\Omega}_i - \boldsymbol{\Omega}_{i-1}||_F + 2\lambda_2 \sum_{i=1}^n |\boldsymbol{\Omega}_i|_1. \tag{4.6}$$

In order to optimize (4.5) using the smoothing technique described in §2.1, we need to show that the gradient of the log-likelihood is Lipschitz continuous. The following Lemma establishes the desired result.

Lemma 6. *The function* $f(\mathbf{A}) = \text{tr}\, \mathbf{S}\mathbf{A} - \log |\mathbf{A}|$ *has Lipschitz continuous gradient on the set* $\{\mathbf{A} \in \mathcal{S}^p \ : \ \Lambda_{\min}(\mathbf{A}) \geq \gamma\}$*, with Lipschitz constant* $L = \gamma^{-2}$.

Following [3], we can show that a solution to the optimization problem (4.5), on each estimated block, is indeed positive definite matrix with smallest eigenvalue bounded away from zero. This allows us to use the Nesterov's smoothing technique to solve (4.5).

Penalized maximum likelihood approach for estimating sparse precision matrix was proposed by [42]. Here, we have modified the penalty to perform estimation under the model (1.1). Although the parameters of the precision matrix can be estimated consistently using the penalized maximum likelihood approach, a number of theoretical results have shown that the neighborhood selection procedure requires lest stringent assumptions in order to estimate the underlying network consistently [28, 32]. We observe this phenomena in our simulation studies as well.

5. Numerical studies

In this section, we present a small numerical study on simulated networks. A full performance test and application on real world data is beyond the scope of this paper which mainly focuses on the theory of time-varying model estimation. In all of our simulations studies we set $p = 30$ and $B = 3$ with $|\mathcal{B}_1| = 80$, $|\mathcal{B}_2| = 130$ and $|\mathcal{B}_3| = 90$, so that in total we have $n = 300$ samples. We consider two types of random networks: a chain and a nearest neighbor network. We measure the performance of the estimation procedure outlined in §2 on the following metrics: average precision of estimated edges, average recall of estimated edges and average F_1 score which combines the precision and recall score. The precision, recall and F_1 score are respectively defined as

$$\begin{aligned} \mathit{precision} &= \frac{1}{n}\sum_{i\in[n]} \frac{\sum_{a\in[p]}\sum_{b=a+1}^{p} \mathbb{1}\{(a,b)\in \hat{E}_i \ \wedge (a,b)\in E_i\}}{\sum_{a\in[p]}\sum_{b=a+1}^{p} \mathbb{1}\{(a,b)\in \hat{E}_i\}} \\ \mathit{recall} &= \frac{1}{n}\sum_{i\in[n]} \frac{\sum_{a\in[p]}\sum_{b=a+1}^{p} \mathbb{1}\{(a,b)\in \hat{E}_i \ \wedge (a,b)\in E_i\}}{\sum_{a\in[p]}\sum_{b=a+1}^{p} \mathbb{1}\{(a,b)\in E_i\}} \\ F_1 &= \frac{2 * \mathit{precision} * \mathit{recall}}{\mathit{precision} + \mathit{recall}}. \end{aligned}$$

Furthermore, we report results on estimating the partition boundaries using $n^{-1}h(\hat{\mathcal{T}}, \mathcal{T})$, where $h(\hat{\mathcal{T}}, \mathcal{T})$ is defined in (3.1). Results are averaged over 50 simulation runs. We compare the TD-Lasso algorithm introduced in §2.1 against an oracle algorithm which exactly knows the true partition boundaries. In this case, it is only needed to run the algorithm of [28] on each block of the partition independently. We use a BIC criterion to select the tuning parameter for this oracle procedure as described in [31]. Furthermore, we report results using neighborhood selection procedures introduced in §4, which are denoted TD_1-Lasso and TD_∞-Lasso, as well as the penalized maximum likelihood procedure, which is denoted as $\text{LL}_{\max}$. We choose the tuning parameters for the penalized maximum likelihood procedure using the BIC procedure.

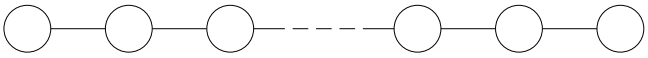

FIG 2. *A chain graph*

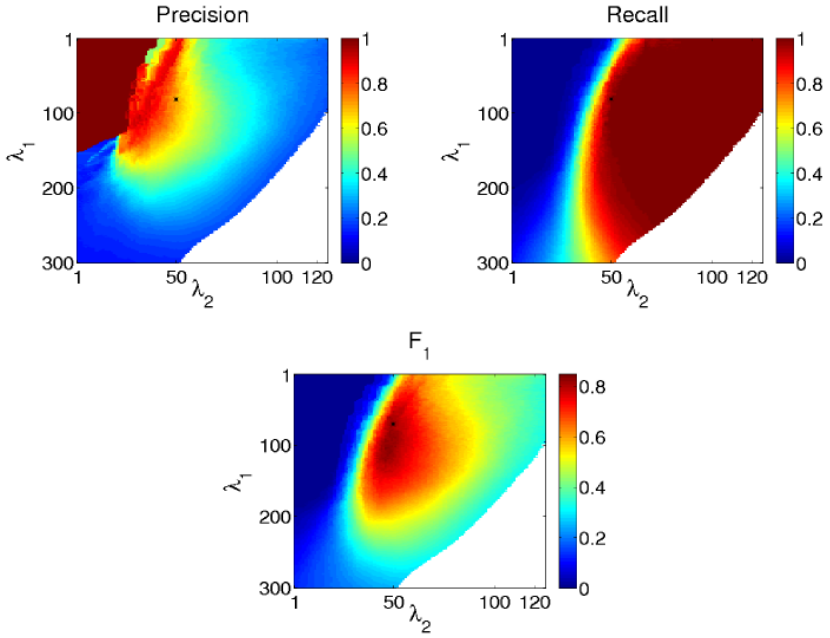

FIG 3. *Plots of the precision, recall and F_1 scores as functions of the penalty parameters λ_1 and λ_2 for chain networks estimated using the TD-Lasso. The parameter λ_1 is obtained as $100 * 0.98^{50+i}$, where i indexes y-axis. The parameter λ_2 is computed as $285 * 0.98^{230+j}$, where j indexes x-axis. Black dot represents the selected tuning parameters. The white region of each plot corresponds to a region of the parameter space that we did not explore.*

Chain networks We follow the simulation in [12] to generate a chain network (see Figure 2). This network corresponds to a tridiagonal precision matrix (after an appropriate permutation of nodes). The network is generated as follows. First, we choose to generate a random permutation π of $[n]$. Next, the covariance matrix is generated as follows: the element at position (a, b) is chosen as $\sigma_{ab} = \exp(-|t_{\pi(a)} - t_{\pi(b)}|/2)$ where $t_1 < t_2 < \cdots < t_p$ and $t_i - t_{i-1} \sim \text{Unif}(0.5, 1)$ for $i = 2, \ldots, p$. This processes is repeated three times to obtain three different covariance matrices, from which we sample 80, 130 and 90 samples respectively.

For illustrative purposes, Figure 3 plots the precision, recall and F_1 score computed for different values of the penalty parameters λ_1 and λ_2. Table 2 shows the precision, recall and F_1 score for the parameters chosen using the BIC score described in 2.2, as well as the error in estimating the partition boundaries. The numbers in parentheses correspond to standard deviation. Due to the fact that there is some error in estimating the partition boundaries,

TABLE 2
Performance of different procedures when estimating chain networks

Method name	Precision	Recall	F_1 score	$n^{-1}h(\hat{\mathcal{T}}, \mathcal{T})$
TD-Lasso	0.84 (0.04)	0.80 (0.04)	0.82 (0.04)	0.03 (0.01)
TD_1-Lasso	0.78 (0.05)	0.70 (0.03)	0.74 (0.04)	N/A
TD_∞-Lasso	0.83 (0.03)	0.80 (0.03)	0.81 (0.03)	0.03 (0.01)
$\text{LL}_{\max}$	0.72 (0.03)	0.65 (0.03)	0.68 (0.04)	0.06 (0.02)
Oracle procedure	0.97 (0.02)	0.89 (0.02)	0.93 (0.02)	0 (0)

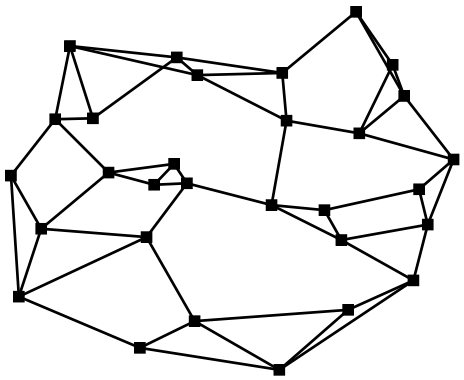

FIG 4. *An instance of a random neighborhood graph with 30 nodes.*

we observe a decrease in performance compared to the oracle procedure that knows the correct position of the partition boundaries. Further, we observe that the neighborhood selection procedure estimate the graph structure more accurately than the maximum likelihood procedure. For TD_1-Lasso we do not report $n^{-1}h(\hat{\mathcal{T}}, \mathcal{T})$, as the procedure does not estimate the partition boundaries.

Nearest neighbors networks We generate nearest neighbor networks following the procedure outlined in [23]. For each node, we draw a point uniformly at random on a unit square and compute the pairwise distances between nodes. Each node is then connected to 4 closest neighbors (see Figure 4). Since some of nodes will have more than 4 adjacent edges, we remove randomly edges from nodes that have degree larger than 4 until the maximum degree of a node in a network is 4. Each edge (a, b) in this network corresponds to a nonzero element in the precision matrix $\mathbf{\Omega}$, whose value is generated uniformly on $[-1, -0.5] \cup [0.5, 1]$. The diagonal elements of the precision matrix are set to a smallest positive number that makes the matrix positive definite. Next, we scale the corresponding covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ to have diagonal elements equal to 1. This processes is repeated three times to obtain three different covariance matrices, from which we sample 80, 130 and 90 samples respectively.

For illustrative purposes, Figure 5 plots the precision, recall and F_1 score computed for different values of the penalty parameters λ_1 and λ_2. Table 3 shows

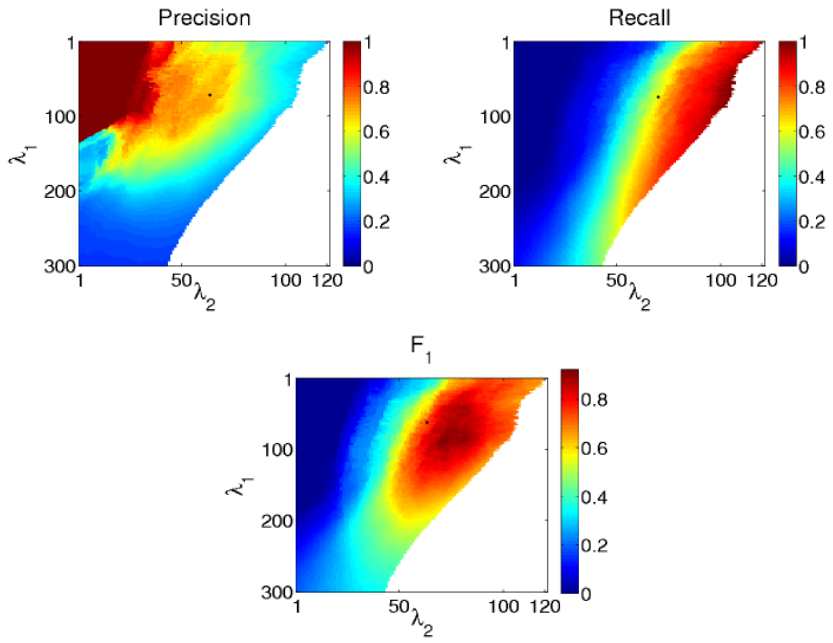

FIG 5. *Plots of the precision, recall and F_1 scores as functions of the penalty parameters λ_1 and λ_2 for nearest neighbor networks estimated using the TD-Lasso. The parameter λ_1 is obtained as $100 * 0.98^{50+i}$, where i indexes y-axis. The parameter λ_2 is computed as $285 * 0.98^{230+j}$, where j indexes x-axis. Black dot represents the selected tuning parameters. The white region of each plot corresponds to a region of the parameter space that we did not explore.*

TABLE 3
Performance of different procedure when estimating random nearest neighbor networks

Method name	Precision	Recall	F_1 score	$n^{-1}h(\hat{\mathcal{T}}, \mathcal{T})$
TD-Lasso	0.79 (0.06)	0.76 (0.05)	0.77 (0.05)	0.04 (0.02)
TD_1-Lasso	0.70 (0.05)	0.68 (0.07)	0.69 (0.06)	N/A
TD_∞-Lasso	0.80 (0.06)	0.75 (0.06)	0.77 (0.06)	0.04 (0.02)
$\text{LL}_{\max}$	0.62 (0.08)	0.60 (0.06)	0.61 (0.06)	0.06 (0.02)
Oracle procedure	0.87 (0.05)	0.82 (0.05)	0.84 (0.04)	0 (0)

the precision, recall, F_1 score and $n^{-1}h(\hat{\mathcal{T}}, \mathcal{T})$ for the parameters chosen using the BIC score, together with their standard deviations. The results obtained for nearest neighbor networks are qualitatively similar to the results obtain for chain networks.

6. Conclusion

We have addressed the problem of time-varying covariance selection when the underlying probability distribution changes abruptly at some unknown 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

and the network structure underlying the sample. The proof technique used to establish the convergence of the boundary fractions using the fused-type penalty is novel and constitutes an important contribution of the paper. 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. The proof technique used to establish the consistency of the network structure builds on the proof for consistency of the neighborhood selection procedure, however, important modifications are necessary since the times of distribution changes are not known in advance. 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. We conjecture that our estimation procedure is also valid in the high-dimensional setting when the number of variables p is much larger than the sample size n. We leave the investigations of the rate of convergence in the high-dimensional setting for a future work.

7. Proofs

7.1. Proof of Lemma 1

For each $i \in [n]$, introduce a $(p-1)$-dimensional vector $\boldsymbol{\gamma}_i$ defined as

$$\boldsymbol{\gamma}_i = \begin{cases} \boldsymbol{\beta}_{\cdot,i} & \text{for } i = 1 \\ \boldsymbol{\beta}_{\cdot,i} - \boldsymbol{\beta}_{\cdot,i-1} & \text{otherwise} \end{cases}$$

and rewrite the objective (2.2) as

$$\{\hat{\boldsymbol{\gamma}}^i\}_{i\in[n]} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{n\times p-1}}{\operatorname{argmin}} \sum_{i=1}^{n}\Bigg(x_{i,a} - \sum_{b\in\setminus a} x_{i,b}\sum_{j\le i}\gamma_{j,b}\Bigg)^2 + 2\lambda_1\sum_{i=2}^{n}||\boldsymbol{\gamma}_i||_2 + 2\lambda_2\sum_{i=1}^{n}\sum_{b\in\setminus a}\Bigg|\sum_{j\le i}\gamma_{j,b}\Bigg|. \tag{7.1}$$

A necessary and sufficient condition for $\{\hat{\boldsymbol{\gamma}}_i\}_{i\in[n]}$ to be a solution of (7.1), is that for each $k \in [n]$ the $(p-1)$-dimensional zero vector, $\mathbf{0}$, belongs to the subdifferential of (7.1) with respect to $\boldsymbol{\gamma}_k$ evaluated at $\{\hat{\boldsymbol{\gamma}}_i\}_{i\in[n]}$, that is,

$$\mathbf{0} = 2\sum_{i=k}^{n}(-\mathbf{x}_{i,\setminus a})\Bigg(x_{i,a} - \sum_{b\in\setminus a} x_{i,b}\hat{\beta}^a_{b,i}\Bigg) + 2\lambda_1\hat{\mathbf{z}}_k + 2\lambda_2\sum_{i=k}^{n}\hat{\mathbf{y}}_i, \tag{7.2}$$

where $\hat{\mathbf{z}}_k \in \partial||\cdot||_2(\hat{\boldsymbol{\gamma}}_k)$, that is,

$$\tilde{\mathbf{z}}_k = \begin{cases} \frac{\tilde{\boldsymbol{\gamma}}_k}{||\tilde{\boldsymbol{\gamma}}_k||_2} & \text{if } \tilde{\boldsymbol{\gamma}}_k \neq 0 \\ \in \mathcal{B}_2(0,1) & \text{otherwise} \end{cases}$$

and for $k \le i$, $\hat{\mathbf{y}}_i \in \partial|\sum_{j\le i}\hat{\boldsymbol{\gamma}}_j|$, that is, $\mathbf{y}_i = \operatorname{sign}(\sum_{j\le i}\hat{\boldsymbol{\gamma}}_j)$ with $\operatorname{sign}(0) \in [-1,1]$. The Lemma now simply follows from (7.2).

7.2. Proof of Theorem 2

We build on the ideas presented in the proof of Proposition 5 in [17]. Using the union bound,

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

and it is enough to show that $\mathbb{P}[|T_j-\tilde{T}_j|>n\delta_n]\to 0$ for all $j\in[B]$. Define the event $A_{n,j}$ as

$$A_{n,j}:=\big\{|T_j-\hat{T}_j|>n\delta_n\big\}$$

and the event C_n as

$$C_n:=\left\{\max_{j\in[B]}|\hat{T}_j-T_j|<\frac{\Delta_{\min}}{2}\right\}.$$

We show that $\mathbb{P}[A_{n,j}]\to 0$ by showing that both $\mathbb{P}[A_{n,j}\cap C_n]\to 0$ and $\mathbb{P}[A_{n,j}\cap C_n^c]\to 0$ as $n\to\infty$. The idea here is that, in some sense, the event C_n is a good event on which the estimated boundary partitions and the true boundary partitions are not too far from each other. Considering the two cases will make the analysis simpler.

First, we show that $\mathbb{P}[A_{n,j}\cap C_n]\to 0$. Without loss of generality, we assume that $\hat{T}_j<T_j$, since the other case follows using the same reasoning. Using (3.2) twice with $k=\hat{T}_j$ and with $k=T_j$ and then applying the triangle inequality we have

$$2\lambda_1\geq\left|\left|\sum_{i=\hat{T}_j}^{T_j-1}\mathbf{x}_{i,\setminus a}\langle\mathbf{x}_{i,\setminus a},\hat{\boldsymbol{\beta}}_{\cdot,i}-\boldsymbol{\beta}_{\cdot,i}\rangle-\sum_{i=\hat{T}_j}^{\hat{T}_j-1}\mathbf{x}_{i,\setminus a}\epsilon_i+\lambda_2\sum_{i=\hat{T}_j}^{T_j-1}\hat{\mathbf{y}}_i\right|\right|_2. \tag{7.3}$$

Some algebra on the above display gives

$$\begin{aligned}2\lambda_1+(T_j-\hat{T}_j)\sqrt{p}\lambda_2&\geq\left|\left|\sum_{i=\hat{T}_j}^{T_j-1}\mathbf{x}_{i,\setminus a}\langle\mathbf{x}_{i,\setminus a},\boldsymbol{\theta}^j-\boldsymbol{\theta}^{j+1}\rangle\right|\right|_2\\&\quad-\left|\left|\sum_{i=\hat{T}_j}^{T_j-1}\mathbf{x}_{i,\setminus a}\langle\mathbf{x}_{i,\setminus a},\boldsymbol{\theta}^{j+1}-\hat{\boldsymbol{\theta}}^{j+1}\rangle\right|\right|_2-\left|\left|\sum_{i=\hat{T}_j}^{T_j-1}\mathbf{x}_{i,\setminus a}\epsilon_i\right|\right|_2\\&=:||R_1||_2-||R_2||_2-||R_3||_2.\end{aligned}$$

The above display occurs with probability one, so that the event $\{2\lambda_1+(T_j-\hat{T}_j)\sqrt{p}\lambda_2\geq\frac{1}{3}||R_1||_2\}\cup\{||R_2||_2\geq\frac{1}{3}||R_1||_2\}\cup\{||R_3||_2\geq\frac{1}{3}||R_1||_2\}$ also occurs with probability one, which gives us the following bound

$$\begin{aligned}\mathbb{P}[A_{n,j} \cap C_n] &\leq \mathbb{P}[A_{n,j} \cap C_n \cap \{2\lambda_1 + (T_j - \hat{T}_j)\sqrt{p}\lambda_2 \geq \frac{1}{3}||R_1||_2\}] \\ &+ \mathbb{P}[A_{n,j} \cap C_n \cap \{||R_2||_2 \geq \frac{1}{3}||R_1||_2\}] \\ &+ \mathbb{P}[A_{n,j} \cap C_n \cap \{||R_3||_2 \geq \frac{1}{3}||R_1||_2\}] \\ &=: \mathbb{P}[A_{n,j,1}] + \mathbb{P}[A_{n,j,2}] + \mathbb{P}[A_{n,j,3}].\end{aligned}$$

First, we focus on the event $A_{n,j,1}$. Using lemma 9, we can upper bound $\mathbb{P}[A_{n,j,1}]$ with

$$\mathbb{P}[2\lambda_1 + (T_j - \hat{T}_j)\sqrt{p}\lambda_2 \geq \frac{\phi_{\min}}{27}(T_j - \hat{T}_j)\xi_{\min}] + 2\exp(-n\delta_n/2 + 2\log n).$$

Since under the assumptions of the theorem $(n\delta_n\xi_{\min})^{-1}\lambda_1 \to 0$ and $\xi_{\min}^{-1}\sqrt{p}\lambda_2 \to 0$ as $n \to \infty$, we have that $\mathbb{P}[A_{n,j,1}] \to 0$ as $n \to \infty$.

Next, we show that the probability of the event $A_{n,j,2}$ converges to zero. Let $\bar{T}_j := \lfloor 2^{-1}(T_j + T_{j+1})\rfloor$. Observe that on the event C_n, $\hat{T}_{j+1} > \bar{T}_j$ so that $\hat{\boldsymbol{\beta}}_{\cdot,i} = \hat{\boldsymbol{\theta}}^{j+1}$ for all $i \in [T_j, \bar{T}_j]$. Using (3.2) with $k = T_j$ and $k = \bar{T}_j$ we have that

$$2\lambda_1 + (\bar{T}_j - T_j)\sqrt{p}\lambda_2 \geq \left|\left| \sum_{i=T_j}^{\bar{T}_j - 1} \mathbf{x}_{i,\setminus a}\langle \mathbf{x}_{i,\setminus a}, \boldsymbol{\theta}^{j+1} - \hat{\boldsymbol{\theta}}^{j+1}\rangle \right|\right|_2 - \left|\left| \sum_{i=T_j}^{\bar{T}_j - 1} \mathbf{x}_{i,\setminus a}\epsilon_i \right|\right|_2.$$

Using lemma 9 on the display above we have

$$||\boldsymbol{\theta}^{j+1} - \hat{\boldsymbol{\theta}}^{j+1}||_2 \leq \frac{36\lambda_1 + 18(\bar{T}_j - T_j)\sqrt{p}\lambda_2 + 18||\sum_{i=T_j}^{\bar{T}_j - 1} \mathbf{x}_{i,\setminus a}\epsilon_i||_2}{(T_{j+1} - T_j)\phi_{\min}}, \tag{7.4}$$

which holds with probability at least $1 - 2\exp(-\Delta_{\min}/4 + 2\log n)$. We will use the above bound to deal with the event $\{||R_2||_2 \geq \frac{1}{3}||R_1||_2\}$. Using lemma 9, we have that $\phi_{\min}(T_j - \hat{T}_j)\xi_{\min}/9 \leq ||R_1||_2$ and $||R_2||_2 \leq (T_j - \hat{T}_j)9\phi_{\max}||\boldsymbol{\theta}^{j+1} - \hat{\boldsymbol{\theta}}^{j+1}||_2$ with probability at least $1 - 4\exp(-n\delta_n/2 + 2\log n)$. Combining with (7.4), the probability $\mathbb{P}[A_{n,j,2}]$ is upper bounded by

$$\begin{aligned}&\mathbb{P}[c_1\phi_{\min}^2\phi_{\max}^{-1}\Delta_{\min}\xi_{\min} \leq \lambda_1] + \mathbb{P}[c_2\phi_{\min}^2\phi_{\max}^{-1}\xi_{\min} \leq \sqrt{p}\lambda_2] \\ &+ \mathbb{P}\left[c_3\phi_{\min}^2\phi_{\max}^{-1}\xi_{\min} \leq (\bar{T}_j - T_j)^{-1}\left|\left| \sum_{i=T_j}^{\bar{T}_j - 1} \mathbf{x}_{i,\setminus a}\epsilon_i \right|\right|_2\right] + c_4\exp(-n\delta_n/2 + 2\log n).\end{aligned}$$

Under the conditions of the theorem, the first term above converges to zero, since $\Delta_{\min} > n\delta_n$ and $(n\delta_n\xi_{\min})^{-1}\lambda_1 \to 0$. The second term also converges to zero, since $\xi_{\min}^{-1}\sqrt{p}\lambda_2 \to 0$. Using lemma 8, the third term converges to zero with the rate $\exp(-c_6\log n)$, since $(\xi_{\min}\sqrt{\Delta_{\min}})^{-1}\sqrt{p\log n} \to 0$. Combining all the bounds, we have that $\mathbb{P}[A_{n,j,2}] \to 0$ as $n \to \infty$.

Finally, we upper bound the probability of the event $A_{n,j,3}$. As before, $\phi_{\min}(T_j - \hat{T}_j)\xi_{\min}/9 \leq ||R_1||_2$ with probability at least $1 - 2\exp(-n\delta_n/2 + 2\log n)$. This

gives us an upper bound on $\mathbb{P}[A_{n,j,3}]$ as

$$\mathbb{P}\left[\frac{\phi_{\min}\xi_{\min}}{27} \leq \frac{||\sum_{i=\hat{T}_j}^{T_j-1}\mathbf{x}_{i,\setminus a}\epsilon_i||_2}{T_j-\hat{T}_j}\right] + 2\exp(-n\delta_n/2+2\log n),$$

which, using lemma 8, converges to zero as under the conditions of the theorem $(\xi_{\min}\sqrt{n\delta_n})^{-1}\sqrt{p\log n}\to 0$. Thus we have shown that $\mathbb{P}[A_{n,j,3}]\to 0$. Since the case when $\hat{T}_j > T_j$ is shown similarly, we have proved that $\mathbb{P}[A_{n,j}\cap C_n]\to 0$ as $n\to\infty$.

We proceed to show that $\mathbb{P}[A_{n,j}\cap C_n^c]\to 0$ as $n\to\infty$. Recall that $C_n^c = \{\max_{j\in[B]}|\hat{T}_j - T_j| \geq \Delta_{\min}/2\}$. Define the following events

$$\begin{aligned} D_n^{(l)} &:= \left\{\exists j\in[B],\ \hat{T}_j \leq T_{j-1}\right\}\cap C_n^c,\\ D_n^{(m)} &:= \left\{\forall j\in[B],\ T_{j-1} < \hat{T}_j < T_{j+1}\right\}\cap C_n^c,\\ D_n^{(r)} &:= \left\{\exists j\in[B],\ \hat{T}_j \geq T_{j+1}\right\}\cap C_n^c \end{aligned}$$

and write $\mathbb{P}[A_{n,j}\cap C_n^c] = \mathbb{P}[A_{n,j}\cap D_n^{(l)}] + \mathbb{P}[A_{n,j}\cap D_n^{(m)}] + \mathbb{P}[A_{n,j}\cap D_n^{(r)}]$. First, consider the event $A_{n,j}\cap D_n^{(m)}$ under the assumption that $\hat{T}_j \leq T_j$. Due to symmetry, the other case will follow in a similar way. Observe that

$$\begin{aligned} &\mathbb{P}[A_{n,j}\cap D_n^{(m)}]\\ &\leq \mathbb{P}[A_{n,j}\cap\{(\hat{T}_{j+1}-T_j)\geq\frac{\Delta_{\min}}{2}\}\cap D_n^{(m)}]\\ &\quad + \mathbb{P}[\{(T_{j+1}-\hat{T}_{j+1})\geq\frac{\Delta_{\min}}{2}\}\cap D_n^{(m)}]\\ &\leq \mathbb{P}[A_{n,j}\cap\{(\hat{T}_{j+1}-T_j)\geq\frac{\Delta_{\min}}{2}\}\cap D_n^{(m)}]\\ &\quad + \sum_{k=j+1}^{B-1}\mathbb{P}[\{(T_k-\hat{T}_k)\geq\frac{\Delta_{\min}}{2}\}\cap\{(\hat{T}_{k+1}-T_k)\geq\frac{\Delta_{\min}}{2}\}\cap D_n^{(m)}]. \end{aligned} \tag{7.5}$$

We bound the first term in (7.5) and note that the other terms can be bounded in the same way. The following analysis is performed on the event $A_{n,j}\cap\{(\hat{T}_{j+1}-T_j)\geq\Delta_{\min}/2\}\cap D_n^{(m)}$. Using (3.2) with $k=\hat{T}_j$ and $k=T_j$, after some algebra (similar to the derivation of (7.3)) the following holds

$$||\boldsymbol{\theta}^j - \hat{\boldsymbol{\theta}}^{j+1}||_2 \leq \frac{18\lambda_1 + 9(T_j-\hat{T}_j)\sqrt{p}\lambda_2 + 9||\sum_{i=\hat{T}_j}^{T_j-1}\mathbf{x}_{i,\setminus a}\epsilon_i||}{\phi_{\min}(T_j-\hat{T}_j)},$$

with probability at least $1-2\exp(-n\delta_n/2+2\log n)$, where we have used lemma 9. Let $\bar{T}_j = \lfloor 2^{-1}(T_j+T_{j+1})\rfloor$. Using (3.2) with $k=\bar{T}_j$ and $k=T_j$ after some al-

gebra (similar to the derivation of (7.4)) we obtain the following bound

$$||\boldsymbol{\theta}^j - \boldsymbol{\theta}^{j+1}||_2 \leq \frac{18\lambda_1 + 9(\bar{T}_j - T_j)\sqrt{p}\lambda_2 + 9||\sum_{i=T_j}^{\bar{T}_j - 1} \mathbf{x}_{i,\setminus a}\epsilon_i||_2}{\phi_{\min}(\bar{T}_j - T_j)} + 81\phi_{\max}\phi_{\min}^{-1}||\boldsymbol{\theta}^j - \hat{\boldsymbol{\theta}}^{j+1}||_2,$$

which holds with probability at least $1 - c_1 \exp(-n\delta_n/2 + 2\log n)$, where we have used lemma 9 twice. Combining the last two displays, we can upper bound the first term in (7.5) with

$$\mathbb{P}[\xi_{\min} n\delta_n \leq c_1\lambda_1] + \mathbb{P}[\xi_{\min} \leq c_2\sqrt{p}\lambda_2] + \mathbb{P}[\xi_{\min}\sqrt{n\delta_n} \leq c_3\sqrt{p\log n}] + c_4\exp(-c_5\log n),$$

where we have used lemma 8 to obtain the third term. Under the conditions of the theorem, all terms converge to zero. Reasoning similar about the other terms in (7.5), we can conclude that $\mathbb{P}[A_{n,j} \cap D_n^{(m)}] \to 0$ as $n \to \infty$.

Next, we bound the probability of the event $A_{n,j} \cap D_n^{(l)}$, which is upper bounded by

$$\mathbb{P}[D_n^{(l)}] \leq \sum_{j=1}^{B} 2^{j-1}\mathbb{P}[\max\{l \in [B] : \hat{T}_l \leq T_{l-1}\} = j].$$

Observe that

$$\{\max\{l \in [B] : \hat{T}_l \leq T_{l-1}\} = j\} \subseteq \bigcup_{l=j}^{B}\{T_j - \hat{T}_j \geq \frac{\Delta_{\min}}{2}\} \cap \{\hat{T}_{j+1} - T_j \geq \frac{\Delta_{\min}}{2}\}$$

so that we have

$$\mathbb{P}[D_n^{(l)}] \leq 2^{B-1}\sum_{j=1}^{B-1}\sum_{l>j}\mathbb{P}[\{T_l - \hat{T}_l \geq \frac{\Delta_{\min}}{2}\} \cap \{\hat{T}_{l+1} - T_l \geq \frac{\Delta_{\min}}{2}\}].$$

Using the same arguments as those used to bound terms in (7.5), we have that $\mathbb{P}[D_n^{(l)}] \to 0$ as $n \to \infty$ under the conditions of the theorem. Similarly, we can show that the term $\mathbb{P}[D_n^{(r)}] \to 0$ as $n \to \infty$. Thus, we have shown that $\mathbb{P}[A_{n,j} \cap C_n^c] \to 0$, which concludes the proof.

7.3. Proof of Lemma 4

Consider $\hat{\mathcal{T}}$ fixed. The lemma is a simple consequence of the duality theory, which states that given the subdifferential $\hat{\mathbf{y}}_i$ (which is constant for all $i \in \hat{\mathcal{B}}^j$, $\hat{\mathcal{B}}^j$ being an estimated block of the partition $\hat{\mathcal{T}}$), all solutions $\{\check{\boldsymbol{\beta}}_{\cdot,i}\}_{i\in[n]}$ of (2.2) need to satisfy the complementary slackness condition $\sum_{b\in\setminus a}\hat{y}_{i,b}\check{\beta}_{b,i} = ||\check{\boldsymbol{\beta}}_{\cdot,i}||_1$, which holds only if $\check{\beta}_{b,i} = 0$ for all $b \in \setminus a$ for which $|\hat{y}_{i,b}| < 1$.

7.4. Proof of Theorem 5

Since the assumptions of theorem 2 are satisfied, we are going to work on the event

$$\mathcal{E} := \{\max_{j\in[B]} |\hat{T}_j - T_j| \leq n\delta_n\}.$$

In this case, $|\hat{\mathcal{B}}^k| = \mathcal{O}(n)$. For $i \in \hat{\mathcal{B}}^k$, we write

$$x_{i,a} = \sum_{b\in S^j} x_{i,b}\theta^k_b + e_i + \epsilon_i \tag{7.6}$$

where $e_i = \sum_{b\in S} x_{i,b}(\beta_{b,i} - \theta^k_b)$ is the bias. Observe that $\forall i \in \hat{\mathcal{B}}^k \cap \mathcal{B}^k$, the bias $e_i = 0$, while for $i \notin \hat{\mathcal{B}}^k \cap \mathcal{B}^k$, the bias e_i is normally distributed with variance bounded by $M^2\phi_{\max}$ under the assumption **A1** and **A3**.

We proceed to show that $S(\hat{\boldsymbol{\theta}}^k) \subset S^k$. Since $\hat{\boldsymbol{\theta}}^k$ is an optimal solution of (2.2), it needs to satisfy

$$\begin{aligned}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{\setminus a})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{\setminus a}(\hat{\boldsymbol{\theta}}^k - \boldsymbol{\theta}^k) - (\mathbf{X}^{\hat{\mathcal{B}}^k}_{\setminus a})'(\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}) \\ + \lambda_1(\hat{\mathbf{z}}_{\hat{T}_{k-1}} - \hat{\mathbf{z}}_{\hat{T}_k}) + \lambda_2|\hat{\mathcal{B}}^k|\hat{\mathbf{y}}_{\hat{T}_{k-1}} = 0.\end{aligned} \tag{7.7}$$

Now, we will construct the vectors $\check{\boldsymbol{\theta}}^k, \check{\mathbf{z}}_{\hat{T}_{k-1}}, \check{\mathbf{z}}_{\hat{T}_k}$ and $\check{\mathbf{y}}_{\hat{T}_{k-1}}$ that satisfy (7.7) and verify that the subdifferential vectors are dual feasible. Consider the following restricted optimization problem

$$\begin{aligned}\min_{\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{\hat{B}};\ \boldsymbol{\theta}^k_{N^k}=\mathbf{0}} \sum_{j\in[\hat{B}]} ||\mathbf{X}^{\hat{\mathcal{B}}^j}_a - \mathbf{X}^{\hat{\mathcal{B}}^j}_{\setminus a}\boldsymbol{\theta}^j||_2^2 \\ + 2\lambda_1\sum_{j=2}^{\hat{B}} ||\boldsymbol{\theta}^j - \boldsymbol{\theta}^{j-1}||_2 + 2\lambda_2\sum_{j=1}^{\hat{B}} |\hat{\mathcal{B}}^j|||\boldsymbol{\theta}^j||_1,\end{aligned} \tag{7.8}$$

where the vector $\boldsymbol{\theta}^k_{N^k}$ is constrained to be $\mathbf{0}$. Let $\{\check{\boldsymbol{\theta}}^j\}_{j\in[\hat{B}]}$ be a solution to the restricted optimization problem (7.8). Set the subgradient vectors as $\check{\mathbf{z}}_{\hat{T}_{k-1}} \in \partial||\check{\boldsymbol{\theta}}^k - \check{\boldsymbol{\theta}}^{k-1}||$, $\check{\mathbf{z}}_{T_k} \in \partial||\check{\boldsymbol{\theta}}^{k+1} - \check{\boldsymbol{\theta}}^k||$ and $\check{\mathbf{y}}_{\hat{T}_{k-1},S^k} = \text{sign}(\check{\boldsymbol{\theta}}^k_{S^k})$. Solve (7.7) for $\check{\mathbf{y}}_{\hat{T}_{k-1},N^k}$. By construction, the vectors $\check{\boldsymbol{\theta}}^k, \check{\mathbf{z}}_{\hat{T}_{k-1}}, \check{\mathbf{z}}_{\hat{T}_k}$ and $\check{\mathbf{y}}_{\hat{T}_{k-1}}$ satisfy (7.7). Furthermore, the vectors $\check{\mathbf{z}}_{\hat{T}_{k-1}}$ and $\check{\mathbf{z}}_{\hat{T}_k}$ are elements of the subdifferential, and hence dual feasible. To show that $\check{\boldsymbol{\theta}}^k$ is also a solution to (3.4), we need to show that $||\check{\mathbf{y}}_{\hat{T}_{k-1},N^k}||_\infty \leq 1$, that is, that $\check{\mathbf{y}}^{\hat{T}_{k-1}}$ is also dual feasible variable. Using lemma 4, if we show that $\check{\mathbf{y}}_{\hat{T}_{k-1},N^k}$ is strict dual feasible, $||\check{\mathbf{y}}_{\hat{T}_{k-1},N^k}||_\infty < 1$, then any other solution $\hat{\check{\boldsymbol{\theta}}}^k$ to (3.4) will satisfy $\hat{\check{\boldsymbol{\theta}}}^k_N = \mathbf{0}$.

From (7.7) we can obtain an explicit formula for $\check{\boldsymbol{\theta}}_{S^k}$

$$\begin{aligned}\check{\boldsymbol{\theta}}^k_{S^k} = \boldsymbol{\theta}^k_{S^k} + \left((\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1}(\mathbf{X}^{\hat{\mathcal{B}}}_{S^k})'(\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}) \\ - \left((\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1}\left(\lambda_1(\check{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \check{\mathbf{z}}_{\hat{T}_k,S^k}) + \lambda_2|\hat{\mathcal{B}}^k|\check{\mathbf{y}}_{\hat{T}_{k-1},S^k}\right).\end{aligned} \tag{7.9}$$

Recall that for large enough n we have that $|\hat{\mathcal{B}}| > p$, so that the matrix $(\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})'\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}$ is invertible with probability one. Plugging (7.9) into (7.7), we have that $||\check{\mathbf{y}}_{\hat{T}_{k-1},N^k}||_\infty < 1$ if $\max_{b\in N^k}|Y_b| < 1$, where Y_b is defined to be

$$Y_b := \left(\mathbf{X}_b^{\hat{\mathcal{B}}^k}\right)'\left[\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})'\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\right)^{-1}\left(\check{\mathbf{y}}_{\hat{T}_{k-1},S^k} + \frac{\lambda_1(\hat{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \hat{\mathbf{z}}_{\hat{T}_k,S^k})}{|\hat{\mathcal{B}}^k|\lambda_2}\right) + \mathbf{H}_{S^k}^{\hat{\mathcal{B}}^k,\perp}\left(\frac{\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}}{|\hat{\mathcal{B}}^k|\lambda_2}\right)\right] - \frac{\lambda_1(\check{z}_{\hat{T}_{k-1},b} - \check{z}_{\hat{T}_k,b})}{|\hat{\mathcal{B}}^k|\lambda_2}, \tag{7.10}$$

where $\mathbf{H}_{S^k}^{\hat{\mathcal{B}}^k,\perp}$ is the projection matrix

$$\mathbf{H}_{S^k}^{\hat{\mathcal{B}}^k,\perp} = \mathbf{I} - \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})'\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\right)^{-1}\left(\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\right)'.$$

Let $\tilde{\boldsymbol{\Sigma}}^k$ and $\hat{\tilde{\boldsymbol{\Sigma}}}^k$ be defined as

$$\tilde{\boldsymbol{\Sigma}}^k = \frac{1}{|\hat{\mathcal{B}}^k|}\sum_{i\in\hat{\mathcal{B}}^k}\mathbb{E}[\mathbf{x}_{\setminus a}^i(\mathbf{x}_{\setminus a}^i)'] \quad \text{and} \quad \hat{\tilde{\boldsymbol{\Sigma}}}^k = \frac{1}{|\hat{\mathcal{B}}^k|}\sum_{i\in\hat{\mathcal{B}}^k}\mathbf{x}_{\setminus a}^i(\mathbf{x}_{\setminus a}^i)'.$$

For $i \in [n]$, we let $\mathcal{B}(i)$ index the block to which the sample i belongs to. Now, for any $b \in N^k$, we can write $x_b^i = \boldsymbol{\Sigma}_{bS^k}^{\mathcal{B}(i)}(\boldsymbol{\Sigma}_{S^kS^k}^{\mathcal{B}(i)})^{-1}\mathbf{x}_{S^k}^i + w_b^i$ where w_b^i is normally distributed with variance $\sigma_b^2 < 1$ and independent of $\mathbf{x}_{S^k}^i$. Let $\mathbf{F}_b \in \mathbb{R}^{|\hat{\mathcal{B}}^k|}$ be the vector whose components are equal to $\boldsymbol{\Sigma}_{bS^k}^{\mathcal{B}(i)}(\boldsymbol{\Sigma}_{S^kS^k}^{\mathcal{B}(i)})^{-1}\mathbf{x}_{S^k}^i$, $i \in \hat{\mathcal{B}}^k$, and $\mathbf{W}_b \in \mathbb{R}^{|\hat{\mathcal{B}}^k|}$ be the vector with components equal to w_b^i. Using this notation, we write $Y_b = T_b^1 + T_b^2 + T_b^3 + T_b^4$ where

$$T_b^1 = \mathbf{F}_b'\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})'\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\right)^{-1}\left(\check{\mathbf{y}}_{\hat{T}_{k-1}} + \frac{\lambda_1(\check{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \check{\mathbf{z}}_{\hat{T}_k,S^k})}{|\hat{\mathcal{B}}^k|\lambda_2}\right) \tag{7.11}$$

$$T_b^2 = \mathbf{F}_b'\mathbf{H}_{S^k}^{\hat{\mathcal{B}}^k,\perp}\left(\frac{\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}}{|\hat{\mathcal{B}}^k|\lambda_2}\right) \tag{7.12}$$

$$T_b^3 = \left(\tilde{\mathbf{W}}_b\right)'\left[\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})'\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k}\right)^{-1}\left(\check{\mathbf{y}}_{\hat{T}_{k-1}} + \frac{\lambda_1(\check{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \check{\mathbf{z}}_{\hat{T}_k,S^k})}{|\hat{\mathcal{B}}^k|\lambda_2}\right) + \mathbf{H}_{S^k}^{\hat{\mathcal{B}}^k,\perp}\left(\frac{\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}}{|\hat{\mathcal{B}}^k|\lambda_2}\right)\right] \tag{7.13}$$

and

$$T_b^4 = -\frac{\lambda_1(\check{z}_{\hat{T}_{k-1},b} - \check{z}_{\hat{T}_k,b})}{|\hat{\mathcal{B}}^k|\lambda_2}. \tag{7.14}$$

We analyze each of the terms separately. Starting with the term T_b^1, after some algebra, we obtain that

$$\begin{aligned}
&\mathbf{F}_b' \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k} \left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})' \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k} \right)^{-1} \\
&\quad = \sum_{j \,:\, \hat{\mathcal{B}}^k \cap \mathcal{B}^j \neq \emptyset} \frac{|\mathcal{B}^j \cap \hat{\mathcal{B}}^k|}{|\hat{\mathcal{B}}^k|} \mathbf{\Sigma}_{bS^k}^j (\mathbf{\Sigma}_{S^k S^k}^j)^{-1} (\hat{\mathbf{\Sigma}}_{S^k S^k}^{\mathcal{B}^j \cap \hat{\mathcal{B}}^k} - \mathbf{\Sigma}_{S^k S^k}^j) \left(\hat{\tilde{\mathbf{\Sigma}}}_{S^k S^k}^{k} \right)^{-1} \\
&\qquad + \tilde{\mathbf{\Sigma}}_{bS^k}^k ((\hat{\tilde{\mathbf{\Sigma}}}_{S^k S^k}^{k})^{-1} - (\tilde{\mathbf{\Sigma}}_{S^k S^k}^k)^{-1}) \\
&\qquad + \tilde{\mathbf{\Sigma}}_{bS^k}^k (\tilde{\mathbf{\Sigma}}_{S^k S^k}^k)^{-1}.
\end{aligned} \tag{7.15}$$

Recall that we are working on the event $\mathcal{E}$, so that $\| \tilde{\mathbf{\Sigma}}_{N^k S^k}^k (\tilde{\mathbf{\Sigma}}_{S^k S^k}^k)^{-1} \|_\infty \xrightarrow{n \to \infty} \| \mathbf{\Sigma}_{N^k S^k}^k (\mathbf{\Sigma}_{S^k S^k}^k)^{-1} \|_\infty$ and $(|\hat{\mathcal{B}}^k| \lambda_2)^{-1} \lambda_1 (\check{\mathbf{z}}_{\hat{T}_{k-1}, S^k} - \check{\mathbf{z}}_{\hat{T}_k, S^k}) \xrightarrow{n \to \infty} 0$ elementwise. Using (7.20) we bound the first two terms in the equation above. We bound the first term by observing that for any j and any $b \in N^k$ and n sufficiently large

$$\begin{aligned}
&\frac{|\mathcal{B}^j \cap \hat{\mathcal{B}}^k|}{|\hat{\mathcal{B}}^k|} ||\mathbf{\Sigma}_{bS^k}^j (\mathbf{\Sigma}_{S^k S^k}^j)^{-1} (\hat{\mathbf{\Sigma}}_{S^k S^k}^{\mathcal{B}^j \cap \hat{\mathcal{B}}^k} - \mathbf{\Sigma}_{S^k S^k}^j)||_\infty \\
&\qquad \leq \frac{|\mathcal{B}^j \cap \hat{\mathcal{B}}^k|}{|\hat{\mathcal{B}}^k|} ||\mathbf{\Sigma}_{bS^k}^j (\mathbf{\Sigma}_{S^k S^k}^j)^{-1}||_1 ||\hat{\mathbf{\Sigma}}_{S^k S^k}^{\mathcal{B}^j \cap \hat{\mathcal{B}}^k} - \mathbf{\Sigma}_{S^k S^k}^j||_\infty \\
&\qquad \leq C_1 \frac{|\mathcal{B}^j \cap \hat{\mathcal{B}}^k|}{|\hat{\mathcal{B}}^k|} ||\hat{\mathbf{\Sigma}}_{S^k S^k}^{\mathcal{B}^j \cap \hat{\mathcal{B}}^k} - \mathbf{\Sigma}_{S^k S^k}^j||_\infty \leq \epsilon_1
\end{aligned}$$

with probability $1 - c_1 \exp(-c_2 \log n)$. Next, for any $b \in N^k$ we bound the second term as

$$\begin{aligned}
&||\tilde{\mathbf{\Sigma}}_{bS^k}^k ((\hat{\tilde{\mathbf{\Sigma}}}_{S^k S^k}^{k})^{-1} - (\tilde{\mathbf{\Sigma}}_{S^k S^k}^k)^{-1})||_1 \\
&\quad \leq C_2 ||(\hat{\tilde{\mathbf{\Sigma}}}_{S^k S^k}^{k})^{-1} - (\tilde{\mathbf{\Sigma}}_{S^k S^k}^k)^{-1})||_F \\
&\quad \leq C_2 ||\tilde{\mathbf{\Sigma}}_{S^k S^k}^k||_F^2 ||\hat{\tilde{\mathbf{\Sigma}}}_{S^k S^k}^k - \tilde{\mathbf{\Sigma}}_{S^k S^k}^k||_F + \mathcal{O}(||\hat{\tilde{\mathbf{\Sigma}}}_{S^k S^k}^k - \tilde{\mathbf{\Sigma}}_{S^k S^k}^k||_F^2) \\
&\leq \epsilon_2
\end{aligned}$$

with probability $1 - c_1 \exp(-c_2 \log n)$. Choosing ϵ_1, ϵ_2 sufficiently small and for n large enough, we have that $\max_b |T_b^1| \leq 1 - \alpha + o_p(1)$ under the assumption **A4**.

We proceed with the term T_b^2, which can be written as

$$\begin{aligned}
T_b^2 = (|\hat{\mathcal{B}}^k| \lambda_2)^{-1} &\left(\mathbf{\Sigma}_{bS^k}^k \left(\mathbf{\Sigma}_{S^k S^k}^k \right)^{-1} - \mathbf{F}_b' \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k} \left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})' \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k} \right)^{-1} \right) \sum_{i \in \mathcal{B}^k \cap \hat{\mathcal{B}}^k} \mathbf{x}_{S^k}^i \epsilon^i \\
+ (|\hat{\mathcal{B}}^k| \lambda_2)^{-1} &\sum_{i \notin \mathcal{B}^k \cap \hat{\mathcal{B}}^k} \left(\mathbf{\Sigma}_{bS^k}^{\mathcal{B}(i)} \left(\mathbf{\Sigma}_{S^k S^k}^{\mathcal{B}(i)} \right)^{-1} - \mathbf{F}_b' \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k} \left((\mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k})' \mathbf{X}_{S^k}^{\hat{\mathcal{B}}^k} \right)^{-1} \right) \mathbf{x}_{S^k}^i (e^i + \epsilon^i).
\end{aligned}$$

Since we are working on the event $\mathcal{E}$ the second term in the above equation is dominated by the first term. Next, using (7.15) together with (7.20), we have

that for all $b \in N^k$

$$||\mathbf{\Sigma}^k_{bS^k}\left(\mathbf{\Sigma}^k_{S^kS^k}\right)^{-1} - \mathbf{F}'_b\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\left((\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1}||_2 = o_p(1).$$

Combining with Lemma 8, we have that under the assumptions of the theorem

$$\max_b |T^2_b| = o_p(1).$$

We deal with the term T^3_b by conditioning on $\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}$ and $\boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}$, we have that $\mathbf{W}_b$ is independent of the terms in the squared bracket in T^3_b, since all $\check{\mathbf{z}}_{\hat{T}_{k-1},S}, \check{\mathbf{z}}_{\hat{T}_k,S}$ and $\hat{\mathbf{y}}_{\hat{T}_{k-1},S}$ are determined from the solution to the restricted optimization problem. To bound the second term, we observe that conditional on $\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}$ and $\boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}$, the variance of T^3_b can be bounded as

$$\begin{aligned}\mathrm{Var}(T^3_b) &\leq ||\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\left((\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1}\check{\eta}_{S^k} + \mathbf{H}^{\hat{\mathcal{B}}^k,\perp}_{S^k}\Big(\frac{\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}}{|\hat{\mathcal{B}}^k|\lambda_2}\Big)||^2_2 \\ &\leq \check{\eta}'_{S^k}\left((\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1}\check{\eta}_{S^k} + \Big\|\frac{\mathbf{e}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}}{|\hat{\mathcal{B}}^k|\lambda_2}\Big\|^2_2,\end{aligned} \tag{7.16}$$

where

$$\check{\eta}_{S^k} = \Big(\check{\mathbf{y}}_{\hat{T}_{k-1},S^k} + \frac{\lambda_1(\check{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \check{\mathbf{z}}_{\hat{T}_k,S})}{|\hat{\mathcal{B}}|\lambda_2}\Big).$$

Using lemma 9 and Young's inequality, the first term in (7.16) is upper bounded by

$$\frac{18}{|\hat{\mathcal{B}}|\phi_{\min}}\left(s + \frac{2\lambda_1^2}{|\hat{\mathcal{B}}|^2\lambda_2^2}\right)$$

with probability at least $1 - 2\exp(-|\hat{\mathcal{B}}^k|/2 + 2\log n)$. Using lemma 7 we have that the second term is upper bounded by

$$\frac{(1+\delta')(1+M^2\phi_{\max})}{|\hat{\mathcal{B}}|\lambda_2^2}$$

with probability at least $1-\exp(-c_1|\hat{\mathcal{B}}^k|\delta'^2+2\log n)$. Combining the two bounds, we have that $\mathrm{Var}(T^3_b) \leq c_1 s(|\hat{\mathcal{B}}^k|)^{-1}$ with high probability, using the fact that $(|\hat{\mathcal{B}}^k|\lambda_2)^{-1}\lambda_1 \to 0$ and $|\hat{\mathcal{B}}^k|\lambda_2 \to \infty$ as $n \to \infty$. Using the bound on the variance of the term T^3_b and the Gaussian tail bound, we have that

$$\max_{b\in N} |T^3_b| = o_p(1).$$

Combining the results, we have that $\max_{b\in N^k}|Y_b| \leq 1 - \alpha + o_p(1)$. For a sufficiently large n, under the conditions of the theorem, we have shown that $\max_{b\in N}|Y_b| < 1$ which implies that $\mathbb{P}[S(\hat{\boldsymbol{\theta}}^k) \subset S^k] \xrightarrow{n\to\infty} 1$.

Next, we proceed to show that $\mathbb{P}[S^k \subset S(\hat{\boldsymbol{\theta}}^k)] \xrightarrow{n\to\infty} 1$. Observe that

$$\mathbb{P}[S^k \not\subset S(\hat{\boldsymbol{\theta}}^k)] \leq \mathbb{P}[||\hat{\boldsymbol{\theta}}^k_{S^k} - \boldsymbol{\theta}^k_{S^k}||_\infty \geq \theta_{\min}].$$

From (7.7) we have that $||\hat{\boldsymbol{\theta}}^k_{S^k} - \boldsymbol{\theta}^k_{S^k}||_\infty$ is upper bounded by

$$\left|\left| \left(\frac{1}{|\hat{\mathcal{B}}^k|}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1} \frac{1}{|\hat{\mathcal{B}}^k|}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'(\tilde{\mathbf{e}}^{\hat{\mathcal{B}}^k} + \boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k})\right|\right|_\infty$$
$$+ \left|\left| \left((\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1} \left(\lambda_1(\check{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \check{\mathbf{z}}_{\hat{T}_k,S^k}) - \lambda_2|\hat{\mathcal{B}}^{\hat{\mathcal{B}}^k}|\check{\mathbf{y}}_{\hat{T}_{k-1},S^k}\right)\right|\right|_\infty.$$

Since $\tilde{e}_i \neq 0$ only on $i \in \hat{\mathcal{B}}^k \backslash \mathcal{B}^k$ and $n\delta_n/|\hat{\mathcal{B}}^k| \to 0$, the term involving $\tilde{\mathbf{e}}^{\hat{\mathcal{B}}^k}$ is stochastically dominated by the term involving $\boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k}$ and can be ignored. Define the following terms

$$T_1 = \left(\frac{1}{|\hat{\mathcal{B}}^k|}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1} \frac{1}{|\hat{\mathcal{B}}^k|}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\boldsymbol{\epsilon}^{\hat{\mathcal{B}}^k},$$
$$T_2 = \left(\frac{1}{|\hat{\mathcal{B}}^k|}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1} \frac{\lambda_1}{|\hat{\mathcal{B}}^k|\lambda_2}(\check{\mathbf{z}}_{\hat{T}_{k-1},S^k} - \check{\mathbf{z}}_{\hat{T}_k,S^k}),$$
$$T_3 = \left(\frac{1}{|\hat{\mathcal{B}}^k|}(\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k})'\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}\right)^{-1} \check{\mathbf{y}}_{\hat{T}_{k-1},S^k}.$$

Conditioning on $\mathbf{X}^{\hat{\mathcal{B}}^k}_{S^k}$, the term T_1 is a $|S^k|$ dimensional Gaussian with variance bounded by c_1/n with probability at least $1 - c_1\exp(-c_2\log n)$ using lemma 9. Combining with the Gaussian tail bound, the term $||T_1||_\infty$ can be upper bounded as

$$\mathbb{P}\left[||T_1||_\infty \geq c_1\sqrt{\frac{\log s}{n}}\right] \leq c_2\exp(-c_3\log n). \tag{7.17}$$

Using lemma 9, we have that with probability greater than $1 - c_1\exp(-c_2\log n)$

$$||T_2||_\infty \leq ||T_2||_2 \leq c_3\frac{\lambda_1}{|\hat{\mathcal{B}}^k|\lambda_2} \to 0$$

under the conditions of theorem. Similarly $||T_3||_\infty \leq c_1\sqrt{s}$, with probability greater than $1 - c_1\exp(-c_2\log n)$. Combining the terms, we have that

$$||\boldsymbol{\theta}^k - \hat{\boldsymbol{\theta}}^k||_\infty \leq c_1\sqrt{\frac{\log s}{n}} + c_2\sqrt{s}\lambda_2$$

with probability at least $1 - c_3\exp(-c_4\log n)$. Since $\theta_{\min} = \Omega(\sqrt{\log(n)/n})$, we have shown that $S^k \subseteq S(\hat{\boldsymbol{\theta}}^k)$. Combining with the first part, it follows that $S(\hat{\boldsymbol{\theta}}^k) = S^k$ with probability tending to one.

7.5. Proof of Lemma 6

We have that $\nabla f(\mathbf{A}) = \mathbf{A}^{-1}$. Then

$$\begin{aligned}||\nabla f(\mathbf{A}) - \nabla f(\mathbf{A}')||_F &= ||\mathbf{A}^{-1} - (\mathbf{A}')^{-1}||_F \\ &\leq \Lambda_{\max}\mathbf{A}^{-1}||\mathbf{A} - \mathbf{A}'||_F\Lambda_{\max}\mathbf{A}^{-1} \\ &\leq \gamma^{-2}||\mathbf{A} - \mathbf{A}'||_F.\end{aligned}$$

Acknowledgments

We are thankful to Zaïd Harchaoui for an early version of his manuscript [17] and many useful discussions. We thank Larry Wasserman and Ankur P. Parikh for providing comments on an early version of this work and many insightful suggestions. Furthermore, we are very grateful to the Associate Editor and two anonymous referees whose suggestions helped to tremendously improve the manuscript.

Appendix

Technical results

In this section we collect some technical results needed for the proves presented in §7.

Lemma 7. *Let $\{\zeta^i\}_{i\in[n]}$ be a sequence of iid $\mathcal{N}(0,1)$ random variables. If $v_n \geq C\log n$, for some constant $C > 16$, then*

$$\mathbb{P}\Big[\bigcap_{\substack{1\leq l<r\leq n\\ r-l>r_n}}\Big\{\sum_{i=l}^{r}(\zeta^i)^2 \leq (1+C)(r-l+1)\Big\}\Big] \geq 1 - \exp(-c_1\log n)$$

for some constant $c_1 > 0$.

Proof. For any $1 \leq l < r \leq n$, with $r - l > v_n$ we have

$$\begin{aligned}\mathbb{P}\Big[\sum_{i=l}^{r}(\zeta^i)^2 \geq (1+C)(r-l+1)\Big] &\leq \exp(-C(r-l+1)/8) \\ &\leq \exp(-C\log n/8)\end{aligned}$$

using (7.21). The lemma follows from an application of the union bound. □

Lemma 8. *Let $\{\mathbf{x}_i\}_{i\in[n]}$ be independent observations from (1.1) and let $\{\epsilon_i\}_{i\in[n]}$ be independent $\mathcal{N}(0,1)$. Assume that* **A1** *holds. If $v_n \geq C\log n$ for some constant $C > 16$, then*

$$\begin{aligned}\mathbb{P}\Big[\bigcap_{j\in[B]}\bigcap_{\substack{l,r\in\mathcal{B}^j\\ r-l>v_n}}\Big\{\frac{1}{r-l+1}\Big|\Big|\sum_{i=l}^{r}\mathbf{x}_i\epsilon_i\Big|\Big|_2 \leq \frac{\phi_{\max}^{1/2}\sqrt{1+C}}{\sqrt{r-l+1}}\sqrt{p(1+C\log n)}\Big\}\Big] \\ \geq 1 - c_1\exp(-c_2\log n),\end{aligned}$$

for some constants $c_1, c_2 > 0$.

Proof. Let $\mathbf{\Sigma}^{1/2}$ denote the symmetric square root of the covariance matrix $\mathbf{\Sigma}_{SS}$ and let $\mathcal{B}(i)$ denote the block $\mathcal{B}^j$ of the true partition such that $i \in \mathcal{B}^j$. With this notation, we can write $\mathbf{x}_i = \left(\mathbf{\Sigma}^{\mathcal{B}(i)}\right)^{1/2} \mathbf{u}_i$ where $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For any $l \leq r \in \mathcal{B}^j$ we have

$$\left|\left|\sum_{i=l}^{r} \mathbf{x}_i \epsilon_i\right|\right|_2 = \left|\left|\sum_{i=l}^{r} \left(\mathbf{\Sigma}^j\right)^{1/2} \mathbf{u}_i \epsilon_i\right|\right|_2 \leq \phi_{\max}^{1/2} \left|\left|\sum_{i=l}^{r} \mathbf{u}_i \epsilon_i\right|\right|_2.$$

Conditioning on $\{\epsilon_i\}_i$, for each $b \in [p]$, $\sum_{i=l}^{r} u_{i,b}\epsilon_i$ is a normal random variable with variance $\sum_{i=l}^{r}(\epsilon_i)^2$. Hence, $||\sum_{i=l}^{r} \mathbf{u}_i\epsilon_i||_2^2/(\sum_{i=l}^{r}(\epsilon_i)^2)$ conditioned on $\{\epsilon_i\}_i$ is distributed according to χ_p^2 and

$$\begin{aligned}\mathbb{P}\left[\frac{1}{r-l+1}\left|\left|\sum_{i=l}^{r} \mathbf{x}_i\epsilon_i\right|\right|_2 \geq \frac{\phi_{\max}^{1/2}\sqrt{\sum_{i=l}^{r}(\epsilon_i)^2}}{r-l+1}\sqrt{p(1+C\log n)} \;\middle|\; \{\epsilon_i\}_{i=l}^{r}\right] \\ \leq \mathbb{P}[\chi_p^2 \geq p(1+C\log n)] \leq \exp(-C\log n/8),\end{aligned}$$

where the last inequality follows from (7.21). Using lemma 7, for all $l, r \in \mathcal{B}^j$ with $r - l > v_n$ the quantity $\sum_{i=l}^{r}(\epsilon_i)^2$ is bounded by $(1+C)(r-l+1)$ with probability at least $1 - \exp(-c_1 \log n)$, which gives us the following bound

$$\begin{aligned}\mathbb{P}\Bigg[\bigcap_{j\in[B]} \bigcap_{\substack{l,r\in\mathcal{B}^j \\ r-l>v_n}} \left\{\frac{1}{r-l+1}\left|\left|\sum_{i=l}^{r}\mathbf{x}_i\epsilon_i\right|\right|_2 \leq \frac{\phi_{\max}^{1/2}\sqrt{1+C}}{\sqrt{r-l+1}}\sqrt{p(1+C\log n)}\right\}\Bigg] \\ \geq 1 - c_1\exp(-c_2\log n).\end{aligned}$$

□

Lemma 9. *Let $\{\mathbf{x}_i\}_{i\in[n]}$ be independent observations from (1.1). Assume that* **A1** *holds. Then for any $v_n > p$,*

$$\mathbb{P}\left[\max_{\substack{1\leq l<r\leq n \\ r-l>v_n}} \Lambda_{\max}\left(\frac{1}{r-l+1}\sum_{i=l}^{r}\mathbf{x}_i\left(\mathbf{x}_i\right)'\right) \geq 9\phi_{\max}\right] \leq 2n^2\exp(-v_n/2)$$

and

$$\mathbb{P}\left[\min_{\substack{1\leq l<r\leq n \\ r-l>v_n}} \Lambda_{\min}\left(\frac{1}{r-l+1}\sum_{i=l}^{r}\mathbf{x}_i\left(\mathbf{x}_i\right)'\right) \leq \phi_{\min}/9\right] \leq 2n^2\exp(-v_n/2).$$

Proof. For any $1 \leq l < r \leq n$, with $r - l \geq v_n$ we have

$$\begin{aligned}\mathbb{P}\left[\Lambda_{\max}\left(\frac{1}{r-l+1}\sum_{i=l}^{r}\mathbf{x}_i\left(\mathbf{x}_i\right)'\right) \geq 9\phi_{\max}\right] &\leq 2\exp(-(r-l+1)/2) \\ &\leq 2\exp(-v_n/2)\end{aligned}$$

using (7.18), convexity of $\Lambda_{\max}(\cdot)$ and **A1**. The lemma follows from an application of the union bound. The other inequality follows using a similar argument. □

Proof of Proposition 3

The following proof follows main ideas already given in theorem 2. We provide only a sketch.

Given an upper bound on the number of partitions $B_{\max}$, we are going to perform the analysis on the event $\{\hat{B} \leq B_{\max}\}$. Since

$$\mathbb{P}[h(\hat{\mathcal{T}}, \mathcal{T}) \geq n\delta_n \mid \{\hat{B} \leq B_{\max}\}] \leq \sum_{B'=B}^{B_{\max}} \mathbb{P}[h(\hat{\mathcal{T}}, \mathcal{T}) \geq n\delta_n \mid \{|\hat{\mathcal{T}}| = B' + 1\}],$$

we are going to focus on $\mathbb{P}[h(\hat{\mathcal{T}}, \mathcal{T}) \geq n\delta_n \mid \{|\hat{\mathcal{T}}| = B' + 1\}]$ for $B' > B$ (for $B' = B$ it follows from theorem 2 that $h(\hat{\mathcal{T}}, \mathcal{T}) < n\delta_n$ with high probability). Let us define the following events

$$\mathcal{E}_{j,1} = \{\exists l \in [B'] \ : \ |\hat{T}_l - T_j| \geq n\delta_n, |\hat{T}_{l+1} - T_j| \geq n\delta_n \text{ and } \hat{T}_l < T_j < \hat{T}_{l+1}\}$$
$$\mathcal{E}_{j,2} = \{\forall l \in [B'] \ : \ |\hat{T}_l - T_j| \geq n\delta_n \text{ and } \hat{T}_l < T_j\}$$
$$\mathcal{E}_{j,3} = \{\forall l \in [B'] \ : \ |\hat{T}_l - T_j| \geq n\delta_n \text{ and } \hat{T}_l > T_j\}.$$

Using the above events, we have the following bound

$$\mathbb{P}[h(\hat{\mathcal{T}}, \mathcal{T}) \geq n\delta_n \mid \{|\hat{\mathcal{T}}| = B' + 1\}] \leq \sum_{j \in [B]} \mathbb{P}[\mathcal{E}_{j,1}] + \mathbb{P}[\mathcal{E}_{j,2}] + \mathbb{P}[\mathcal{E}_{j,3}].$$

The probabilities of the above events can be bounded using the same reasoning as in the proof of theorem 2, by repeatedly using the KKT conditions given in (3.2). In particular, we can use the strategy used to bound the event $A_{n,j,2}$. Since the proof is technical and does not reveal any new insight, we omit the details.

A collection of known results

This section collects some known results that we have used in the paper. We start by collecting some results on the eigenvalues of random matrices. Let $\mathbf{x} \overset{iid}{\sim} \mathcal{N}(0, \mathbf{\Sigma})$, $i \in [n]$, and $\hat{\mathbf{\Sigma}} = n^{-1} \sum \mathbf{x}_i (\mathbf{x}_i)'$ be the empirical covariance matrix. Denote the elements of the covariance matrix $\mathbf{\Sigma}$ as $[\sigma_{ab}]$ and of the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ as $[\hat{\sigma}_{ab}]$.

Using standard results on concentration of spectral norms and eigenvalues [10], [38] derives the following two crude bounds that can be very useful. Under the assumption that $p < n$,

$$\mathbb{P}[\Lambda_{\max}(\hat{\mathbf{\Sigma}}) \geq 9\phi_{\max}] \leq 2\exp(-n/2) \tag{7.18}$$
$$\mathbb{P}[\Lambda_{\min}(\hat{\mathbf{\Sigma}}) \leq \phi_{\min}/9] \leq 2\exp(-n/2). \tag{7.19}$$

From Lemma A.3. in [6] we have the following bound on the elements of the covariance matrix

$$\mathbb{P}[|\hat{\sigma}_{ab} - \sigma_{ab}| \geq \epsilon] \leq c_1 \exp(-c_2 n\epsilon^2), \qquad |\epsilon| \leq \epsilon_0 \tag{7.20}$$

where c_1 and c_2 are positive constants that depend only on $\Lambda_{\max}(\mathbf{\Sigma})$ and ϵ_0.

Next, we use the following tail bound for χ^2 distribution from [25], which holds for all $\epsilon > 0$,

$$\mathbb{P}[\chi_n^2 > n + \epsilon] \leq \exp(-\frac{1}{8}\min(\epsilon, \frac{\epsilon^2}{n})). \tag{7.21}$$

References

[1] A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proc. Natl. A. Sci.*, 106(29):11878–11883, 2009.

[2] J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998. MR1616121

[3] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9: 485–516, 2008. MR2417243

[4] O. Banerjee, L. El Ghaoui, and A. 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. MR2417243

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009. MR2486527

[6] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Stat.*, 36(1):199–227, 2008. MR2387969

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. MR2061575

[8] P. Brucker. An o(n) algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3): 163–166, 1984. MR0761510

[9] F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization. *Electron. J. Stat.*, 2:1153, 2008. MR2461898

[10] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1: 317–366, 2001. MR1863696

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

[12] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Stat.*, 3(2):521–541, 2009. MR2750671

[13] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008.

[14] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007. MR2391486

[15] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint Structure Estimation for Categorical Markov Networks. Technical report, Department of Statistics, University of Michigan, 2010.

[16] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint Estimation of Multiple Graphical Models. *Biometrika, to appear*, 2010. MR2804206

[17] Zaïd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total-variation penalty. *J. Am. Stat. Soc.*, 105(492):1480–1493, 2010. MR2796565

[18] T. Hastie and R. Tibshirani. Varying-coefficient models. *J. R. Stat. Soc. B*, 55(4):757–796, 1993. ISSN 00359246. MR1229881

[19] M. Kolar and E. P. Xing. Sparsistent estimation of Time-Varying discrete markov random fields. Technical report, Machine Learning Department, Carnegie Mellon University, 2009. Available at arxiv 0907.2337.

[20] M. Kolar, A. P. Parikh, and E. P. Xing. On sparse nonparametric conditional covariance selection. In *Proc 27th Ann. Int'l Conf. Machine Learn.*, 2010.

[21] M. Kolar, L. Song, A. Ahmed, and E. P. Xing. Estimating Time-Varying networks. *Ann. Appl. Statist*, 4(1):94—123, 2010. MR2758086

[22] S. L. Lauritzen. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, 1996. MR1419991

[23] H. Li and J. Gui. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302, 2006.

[24] J. Liu, S. Wu, and J. V Zidek. On segmented multivariate regression. *Stat. Sin.*, 7:497–526, 1997. MR1466692

[25] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in Multi-Task learning. In *Proc. Conf. Learning Theory*, 2009.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. MR2591620

[27] E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997. MR1429931

[28] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.*, 34(3):1436–1462, 2006. MR2278363

[29] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005. MR2166537

[30] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76:2007, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

[31] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Ass.*, 104(486):735–746, 2009. MR2541591

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

determinant divergence. Technical report, Department of Statistics, University of California, Berkeley, 2008.

[33] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1 regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319, 2010. MR2662343

[34] A. Rinaldo. Properties and refinements of the fused lasso. *Ann. Stat.*, 37(5):2922–2952, 2009. MR2541451

[35] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008. MR2417391

[36] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. B*, 67(1):91–108, 2005. MR2136641

[37] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. MR2576316

[38] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE T. Inform. Theory*, 55(5):2183–2202, 2009. MR2729873

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

[40] P. Wang, D. L. Chao, and L. Hsu. Learning networks from high dimensional binary data: An application to genomic instability data. *Biometrics, to appear*, 2009. MR2898828

[41] J. Yin, Z. Geng, R. Li, and H. Wang. Nonparametric Covariance Model. *Statistica Sinica*, 20:469–479, 2010. MR2640671

[42] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. MR2367824

[43] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. MR2274449

[44] S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. In *Proc. Conf. Learning Theory*, pages 455–466, 2008.