Scalable Inference of Sparsely-changing Gaussian Markov Random Fields

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of inferring time-varying Gaussian Markov random fields, where the underlying graphical model is both sparse and changes sparsely over time. Most of the existing methods for the inference of time-varying Markov random fields (MRFs) rely on the *regularized maximum likelihood estimation* (MLE), that typically suffer from weak statistical guarantees and high computational time. Instead, we introduce a new class of constrained optimization problems for the inference of sparsely-changing Gaussian MRFs (GMRFs). The proposed optimization problem is formulated based on the exact ℓ_0 regularization, and can be solved in near-linear time and memory. Moreover, we show that the proposed estimator enjoys a provably small estimation error. We derive sharp statistical guarantees in the high-dimensional regime, showing that such problems can be learned with as few as one sample per time period. Our proposed method is extremely efficient in practice: it can accurately estimate sparsely-changing GMRFs with more than 500 million variables in less than one hour.

1 Introduction

Contemporary systems are comprised of massive numbers of interconnected components that interact according to a hierarchy of complex, unknown, and time-varying topologies. For example, with billions of neurons and hundreds of thousands of voxels, the human brain is considered as one of the most complex physiological networks [19, 23, 29, 31, 38]. The temporal behavior of today's interconnected systems can be captured via *time-varying Markov random fields (MRF)*. Time-varying MRFs are associated with a temporal sequence of undirected *Markov graphs* $\mathcal{G}_t(V, E_t)$, where V and E_t are the set of nodes and edges in the graph at time t. The node set V represents the random variables in the model, while the edge set E_t captures the conditional dependency between these variables at time t. A popular approach for the inference of MRFs is based on the *maximum-likelihood estimation* (MLE): to obtain a model based on which the observed data is most probable to occur [43].

Despite being known as theoretically powerful tools [21, 40], MLE-based methods suffer from several fundamental drawbacks which render them impractical in realistic settings. First, they often suffer from notoriously high computational cost in massive problems, where the number of variables to be inferred is in the order of millions, or more. Second, they struggle to incorporate sparsity amongst their components, which is pervasive in large-scale systems. In particular, while *sparsely-changing* MRFs can in theory be accurately estimated using sparsity-promoting regularizers (such as ℓ_0 penalty), most of the existing methods resort to relaxed or weaker variants of such regularization (such as ℓ_1 penalty), thereby suffering from inferior statistical guarantees.

To address the aforementioned challenges, we propose a class of constrained optimization problems that achieve superior statistical and computational guarantees, compared to the regularized MLE, for the inference of time-varying Gaussian MRFs (GMRFs). Our approach departs from the usual

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

wisdom in statistics and machine learning that inference problems with nonconvex ℓ_0 terms are intractable, and convex proxies should be used instead. In particular, we show that the inference of sparsely-changing Gaussian MRFs can be solved efficiently via nonconvex ℓ_0 penalties.

Notations. The ith element of a time-series vector v_t is denoted as $v_{t;i}$; the $(i,j)^{th}$ element of a time-series matrix V_t is denoted as $V_{t;ij}$. For a vector v, the notation $v_{i:j}$ is used to denote the subvector of v from index i to j. For a vector v, the notations $\|v\|_\infty$, $\|v\|_2$, $\|v\|_0$ denote the ℓ_∞ norm, ℓ_2 norm, and the number of nonzero elements, respectively. Moreover, for a matrix M, the notations $\|M\|_2$, $\|M\|_\infty$, $\|M\|_{1/1}$, $\|M\|_{\infty/\infty}$ refer to the induced 2-norm, induced ∞-norm, ℓ_1/ℓ_1 norm, and ℓ_∞/ℓ_∞ norm, respectively. Moreover, we define $\|M\|_{\text{off}} = \|M\|_{1/1} - \sum_{i=1}^d |M_{ii}|$. For a vector v and matrix M, the notations $\text{supp}(v)$ and $\text{supp}(M)$ are defined as the sets of their nonzero elements. Given two sequences $f(n)$ and $g(n)$, the notation $f(n) \lesssim g(n)$ implies that there exists a constant $C < \infty$ that satisfies $f(n) \leq Cg(n)$, and $f(n) \asymp g(n)$ implies that $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$.

All proofs are deferred to the supplementary file.

1.1 Warm-up: Regularized MLE for Sparsely-changing GMRFs

Consider a multivariate zero-mean Gaussian process $\{X_t\}_{t=0}^T$ with distribution

$$\mathbb{P}(X_t) = \exp\left\{-\frac{1}{2}\langle \Theta_t, X_t X_t^\top\rangle + \langle \eta_t, X_t\rangle - A(\mu_t, \Theta_t)\right\} \tag{1}$$

for $t = 0, \ldots, T$ where $A : \mathbb{R}^{d\times d} \to \mathbb{R}$ is the log-partition function used to normalize the distribution. Without loss of generality, we assume that the mean is zero. At any given time t, a sequence of data samples $\{X_t^{(i)}\}_{i=1}^{N_t}$ is collected from (1). The inference of time-varying GMRFs reduces to estimating the time-varying precision matrix Θ_t from the data samples $\{X_t^{(i)}\}_{i=1}^{N_t}$. Moreover, the edge set of the Markov graph $\mathcal{G}_t$ coincides with the off-diagonal nonzero elements of Θ_t [46].

We first illustrate the fundamental drawbacks of the ℓ_1-regularized MLE for time-varying GMRFs with sparsely-changing structures. The sparse precision matrices can be estimated via the following regularized MLE, also known as *time-varying Graphical Lasso (GL)* [12, 17]:

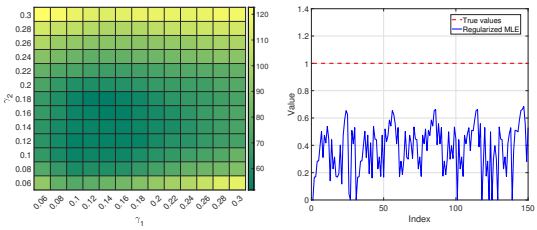
$$\{\widehat{\Theta}_t\}_{t=0}^T = \arg\min_{\Theta_t} \sum_{t=0}^T \Big(\langle \Theta_t, \widehat{\Sigma}_t\rangle - \log\det(\Theta_t)\Big)$$

$$+ \gamma_1 \sum_{t=0}^T \|\Theta_t\|_{\text{off}} + \gamma_2 \sum_{t=1}^T \|\Theta_t - \Theta_{t-1}\|_{1/1} \tag{2a}$$

$$\text{s.t. } \Theta_t \succ 0 \qquad t = 0, 1, \ldots, T \tag{2b}$$

where $\widehat{\Sigma}_t \in \mathbb{R}^{d\times d}$ is the sample covariance matrix at time t. Without loss of generality, we assume that the samples have zero mean. Example 1 below shows that (2) may lead to poor estimates.

Example 1. *Consider a scenario where $\{\Theta_t\}_{t=0}^4 \in \mathbb{R}^{25\times 25}$ are randomly generated symmetric and sparse matrices. At each time $t = 0, \ldots, 4$, the precision matrix Θ_t has exactly 30 off-diagonal elements with value one in its upper-triangular part, and the remaining off-diagonal entries are set to zero. Moreover, the diagonal entries $\Theta_{t;ii}$ are chosen as $1 + \sum_{j\neq i} \Theta_{t;ij}$. At every time, 5 nonzero off-diagonal elements are changed to zero, and 5 zero elements are set to one. The sample covariance $\widehat{\Sigma}_t$ is obtained by collecting 500 samples from the Gaussian distribution with the constructed precision matrices. Figure 1a illustrates a heatmap of the mismatch error, i.e., the total number of mismatches in the sparsity patterns of the true and estimated precision matrices and their differences, for different values of the regularization coefficients. It can be seen that after an exhaustive search over the regularization coefficient space, the best achievable mismatch error is*



(a) (b)

Figure 1: (a) The heatmap of the mismatch error. (b) The true and estimated nonzero elements of the precision matrix.

in the order of 50. Thus, the estimated parameters reveal little information about the true structure of the time-varying GMRF. Moreover, Figure 1b depicts the concatenation of the nonzero elements in the true precision matrices (dashed red line), and their corresponding values in the estimated matrices (blue curve). It can be seen that, even when the sparsity pattern of the elements is correctly recovered, the estimated nonzero entries are "shrunk" toward zero, incurring a substantial bias.

The above example shows the inferior statistical performance of the time-varying GL as an instance of a regularized MLE method for sparsely-changing GMRFs. In addition to its subpar statistical performance, time-varying GL suffers from expensive computational complexity: a general-purpose interior-point algorithm for solving time-varying GL has a prohibitive *per-iteration* complexity of $\mathcal{O}(Td^6)$ [11]. More recent algorithms for solving time-varying GL have lower *per-iteration* complexity of $\mathcal{O}(Td^3)$ [17, 27, 33]. However, these methods suffer from a slow (sublinear) convergence rate of $O(1/\epsilon)$, increasing the overall complexity to $\mathcal{O}(Td^3/\epsilon)$ in order to obtain an ϵ-accurate solution. Thus, there is an inherent tradeoff between the quality of the solution found and the computational time required, with the performance deteriorating sharply as the number of precision digits increase. Solvers with such computational complexity may fall short of practical use in the large-scale settings. We now discuss the proposed method, which finds optimal solutions to the relevant optimization problems in *strongly polynomial time*.

2 Proposed Approach

The proposed framework is based on exact solutions to a class of tractable discrete ℓ_0-problems, thus circumventing bias and other drawbacks of the standard ℓ_1-approximations, while guaranteeing the scalability of the proposed method. As a general framework, we study the optimization problem:

$$\min \ (1-\gamma)\sum_{t=0}^{T}\|\Theta_t\|_0 + \gamma\sum_{t=1}^{T}\|\Theta_t - \Theta_{t-1}\|_0 \tag{3a}$$

$$\text{s.t.} \ \|\Theta_t - \widetilde{F}^*(\widehat{\Sigma}_t)\|_{\infty/\infty} \leq \lambda_t \qquad t = 0, 1, \ldots, T \tag{3b}$$

$$\Theta_t \in \mathbb{R}^{d\times d} \qquad t = 0, 1, \ldots, T \tag{3c}$$

where the optimal solutions $\{\widehat{\Theta}_t\}_{t=0}^T$ are the estimates of the precision matrices of the sparsely-changing GMRF, $\widehat{\Sigma}_t \in \mathbb{R}^{d\times d}$ is the sample covariance matrix at time t, and $\widetilde{F}^*(\cdot)$ is an *approximate backward mapping* of the model. In particular, we use the approximate backward mapping proposed in [47], see §4.1 for a formal definition.

First, we establish a deterministic guarantee on the estimation error of the optimal solution to (3).

Theorem 1 (Estimation error and sparsistency)**.** *Let $0 < \gamma < 1$. For every $t = 0, \ldots, T$, define $\mathcal{S}_t$ as the set of indices corresponding to the nonzero elements of the true precision matrix Θ_t^*, and define $\mathcal{D}_t$ as the set of indices corresponding to the nonzero elements of $\Theta_t^* - \Theta_{t-1}^*$. Assume that*

- $\left\|\Theta_t^* - \widetilde{F}^*(\widehat{\Sigma}_t)\right\|_{\infty/\infty} < \lambda_t$, $\forall 0 \leq t \leq T$,
- $2\lambda_t \leq \min_{(i,j)\in\mathcal{S}_t} |\Theta_{t;ij}^*|$, $\forall 0 \leq t \leq T$,
- $2\lambda_t + 2\lambda_{t-1} \leq \min_{(i,j)\in\mathcal{D}_t} |\Theta_{t;ij}^* - \Theta_{t-1;ij}^*|$, $\forall 0 \leq t \leq T$.

Then, the following statements hold for every $0 \leq t \leq T$:

Sparsistency $\operatorname{supp}\left(\widehat{\Theta}_t\right) = \operatorname{supp}\left(\Theta_t^*\right)$ *and* $\operatorname{supp}\left(\widehat{\Theta}_t - \widehat{\Theta}_{t-1}\right) = \operatorname{supp}\left(\Theta_t^* - \Theta_{t-1}^*\right)$.

Estimation error $\|\widehat{\Theta}_t - \Theta_t^*\|_{\infty/\infty} \leq 2\lambda_t$ *and* $\|\widehat{\Theta}_t - \Theta_t^*\|_2 \leq 2\sqrt{|\mathcal{S}_t|}\lambda_t$.

Theorem 1 presents a set of conditions under which the proposed estimation method achieves sparsistency and small estimation error. The first condition entails that the true precision matrix Θ_t^* is a feasible solution to (3). The second and third conditions imply that there is a non-negligible gap between the zero and nonzero elements of the true parameters and their temporal changes. In §4 we present additional bounds specific to our choice of backward mapping.

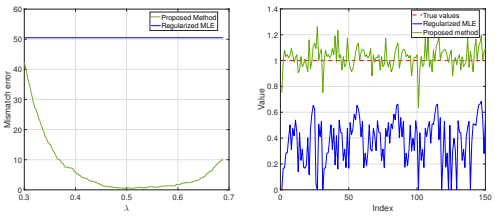
Theorem 2 (Computational complexity)**.** *Given $\widetilde{F}^*(\widehat{\Sigma}_t)$, the optimization problem* (3) *can be solved to optimality in at most $\mathcal{O}((dT)^2)$ time and memory on a single thread.*

Theorem 2 shows that the optimization problem (3) can be solved efficiently and in *strongly polynomial* time despite its non-convex nature. As will be explained later, our choice of approximate

backwards mapping [47] requires inverting a $d \times d$ matrix, thereby increasing the overall complexity to $\mathcal{O}((dT)^2 + d^3)$. If $T \leq \sqrt{d}$, the complexity of the our method is dominated by that of the matrix inversion, which is unavoidable, even if the sample covariance matrix coincides with its true analog.

Our solution method for (3) relies on the *element-wise decomposability* of (3): we decompose (3) into smaller subproblems over different coordinates of $\{\Theta_t\}_{t=0}^T$. Then, we show that the optimal solution to each subproblem can be obtained by solving a shortest path problem on an auxiliary *weighted directed acyclic graph (DAG)*. The details of the solution method are presented in §5. Moreover, the proposed algorithm is easily parallelizable, leading to better runtimes in practice.

Example 1 (continued). *Figure 2 depicts the performance of the proposed method, compared to the regularized MLE with $\gamma_1 = 0.14$ and $\gamma_2 = 0.16$ (corresponding to the smallest mismatch error) for the instances generated in Example 1. The regularization parameter γ in the objective function of* (3) *is set to* 0.2*. Moreover, for simplicity, we set $\lambda_0 = \cdots = \lambda_4 = \lambda$. Figure 2a demonstrates that the proposed method enjoys a significantly smaller mismatch error, for a wide range of λ. On the other hand, Figure 2b shows that the synthetic bias caused by the ℓ_1 penalty in the regularized MLE is alleviated via the proposed method.*



(a) (b)

Figure 2: (a) The mismatch error of the proposed method for different values of λ compared to the regularized MLE. (b) The true and estimated nonzero elements of the precision matrix.

3 Related Works

Time-varying MRF. In addition to the time-varying GL introduced in §1.1, a recent line of works have studied the inference of smoothly-changing GMRFs [24, 16, 49], where a kernel averaging technique combined with Graphical Lasso is used to estimate the smoothly-changing precision matrices. However, these methods do not leverage the prior information about the sparsity of the parameter differences. With the goal of addressing this deficiency, several works have studied the inference of sparsely-changing MRF (also known as sparse *differential networks*) [44, 48, 28]. However, the main drawback of these methods is that they only estimate the parameter differences, and their theoretical guarantees are restricted to problems with two time steps ($T = 0, 1$). Similarly, regression-based approaches have been proposed for change point detection problems [25, 37] with MRFs and two time periods, assuming the sparsity pattern of all entries of the precision matrices change at the same time. In contrast, [45] studies the inference of sparse MRFs given an index variable under the assumption that the sparsity pattern is invariant, whereas [15] assumes that the precision matrix is a linear function of the index variable.

Sparsity-promoting optimization. Optimization problems with ℓ_0 terms are often deemed to be intractable, and approximations are solved instead. Perhaps the most popular approach is the *fused lasso* [35, 41, 39, 42], which calls for replacing terms $\|\Theta_t\|_0$ and $\|\Theta_t - \Theta_{t-1}\|_0$ with their ℓ_1-approximations. Nonetheless, such approximations result in subpar statistical performance when compared with exact ℓ_0 methods [20, 30].

Exact or near-optimal methods for optimization problems of the form

$$\min_{\theta \in [\ell, u]^p} \|\theta\|_0 + \sum_{i=1}^{p} \sum_{j=i+1}^{p} g_{ij}(\theta_i - \theta_j), \tag{4}$$

for given one-dimensional functions $g_{ij} : \mathbb{R} \to \mathbb{R}$, have also been studied in the literature. If functions g_{ij} are convex, then problem (4) admits pseudo-polynomial time algorithm [3, 8]. Moreover, convex relaxations that deliver near-optimal solutions for (4) were proposed for the special case of convex quadratic g functions [7]; if, additionally, we have $\ell = 0$ and $u = \infty$, then problem (4) is in fact solvable in strongly polynomial time [6]. On the other hand, problem (4) is much more challenging for non-convex g: if $g(x) = \mathbb{1}\{x \neq 0\}$, as is the case in (3), then problem (4) is NP-hard even if the term $\|\theta\|_0$ is dropped from the objective [18]. Nonetheless, as we show in this paper, problem (4) can be solved efficiently in the context of time-varying MRFs, where $g_{ij}(x) = 0$ whenever $j > i + 1$.

4 Statistical analysis

Theorem 1 presents a set of deterministic conditions under which the estimates from (3) enjoy zero mismatch error and small estimation error. However, the formulation of (3) is contingent upon the availability of an accurate backward mapping, and a choice of λ_t that satisfies the conditions of Theorem 1. In this section, we show how to efficiently design sample-efficient approximate backward mappings, and select λ_t accordingly for the class of sparsely-changing GMRFs. Moreover, we use the deterministic conditions of Theorem 1 to arrive at a non-asymptotic probabilistic guarantee for the inference of time-varying GMRFs under different prior knowledge on their temporal behavior, such as sparsity and smoothness.

4.1 Sparsely-changing GMRFs

Given the true covariance matrix Σ_t, the backward mapping of time-varying GMRFs as defined in (1) takes the form $F^*(\Sigma_t) = \Sigma_t^{-1}$. In light of this closed-form expression for the backward mapping, a commonly-used approximation is $\widetilde{F}^*(\widehat{\Sigma}_t) = \widehat{\Sigma}_t^{-1}$, where $\widehat{\Sigma}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} X_t^{(i)} X_t^{(i)\top}$ is the sample covariance matrix. However, in the high-dimensional settings where $d \gg N_t$, this approximate backward mapping is not well-defined, since the sample covariance matrix is highly rank-deficient.

To address this issue, [47] propose a *proxy* backward mapping for high-dimensional settings: consider the soft-thresholding operator $\mathtt{ST}_\nu(M) : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$, where $\mathtt{ST}_\nu(M)_{ij} = M_{ij} - \operatorname{sign}(M_{ij})\min\{|M_{ij}|, \nu\}$ if $i \neq j$, and $\mathtt{ST}_\nu(M)_{ij} = M_{ij}$ if $i = j$. The approximate backward mapping is then given by $\widetilde{F}^*(\widehat{\Sigma}_t) = [\mathtt{ST}_\nu(\widehat{\Sigma}_t)]^{-1}$, which is well-defined, even in the high-dimensional setting, with an appropriate choice of the threshold ν [47]. We make the following assumption.

Assumption 1 (Bounded norm)**.** *There exist constant numbers $\kappa_1 < \infty$, $\kappa_2 > 0$, and $\kappa_3 < \infty$ such that, for every $t = 0, \ldots, T$,*

$$\|\Theta_t\|_\infty \leq \kappa_1, \quad \inf_{w:\|w\|_\infty = 1} \|\Sigma_t w\|_\infty \geq \kappa_2, \ \|\Sigma_t\|_{\infty/\infty} \leq \kappa_3.$$

Assumption 1 implies that the true covariance matrices and their inverses have bounded norms. Another key notion which plays an important role in our statistical guarantees is the *weak sparsity* of the covariance matrices, as defined next.

Definition 1 (Weak sparsity)**.** *Given $0 \leq q \leq 1$ and dimension d, define $s_d(q,d) \stackrel{def}{=} \max_i \sum_{j=1}^d |[\Sigma_t]_{ij}|^q$.*

Assuming that $s_d(0,d) \ll d$, Assumption 1 reduces to the covariance matrix being sparse. Moreover, in many cases, a sparse inverse covariance matrix leads to weakly sparse covariance matrices. For instance, if Θ_t has a banded structure with small bandwidth, then it is known that the elements of $\Sigma_t = \Theta_t^{-1}$ enjoy exponential decay away from the main diagonal elements [13, 22]. Under such circumstances, one can verify that $s_d(q,d) \leq \frac{C}{1-\rho^q}$ for some constant $C > 0$ and $\rho < 1$. More generally, a similar statement holds for a class of inverse covariance matrices whose support graphs have large average path length [9, 10]; a large class of inverse covariance matrices with row- and column-sparse structures satisfy this condition. Theorem 3 states that the proposed method results in high-quality solutions provided that the number of samples N_t is sufficiently large with respect to the weak sparsity of the covariance matrices.

Theorem 3. *Consider a sparsely-changing GMRF and let $\zeta = \max\{\log_d(T+1), 1\}$. Given an arbitrary $\tau > \zeta + 2$, let $N_t \gtrsim s_d(q,d)^{\frac{2}{1-q}} \tau \log d$, for some $0 \leq q < 1$. Then, with the approximate backward mapping $\widetilde{F}^*(\widehat{\Sigma}_t) = [\mathcal{ST}_{\nu_t}(\widehat{\Sigma}_t)]^{-1}$, and parameters $\nu_t \asymp \sqrt{\frac{\tau \log d}{N_t}}$ and $\lambda_t \asymp \sqrt{\frac{\tau \log d}{N_t}}$, the estimates $\{\widehat{\Theta}_t\}_{t=0}^T$ obtained from (3) satisfy the following statements for all $t = 0, 1, \ldots, T$, with probability of at least $1 - 4d^{-\tau+\zeta+2}$:*

Sparsistency *We have* $\operatorname{supp}\left(\widehat{\Theta}_t\right) = \operatorname{supp}\left(\Theta_t^*\right)$ *and* $\operatorname{supp}\left(\widehat{\Theta}_t - \widehat{\Theta}_{t-1}\right) = \operatorname{supp}\left(\Theta_t^* - \Theta_{t-1}^*\right)$.

Estimation error $\|\widehat{\Theta}_t - \Theta_t^*\|_{\infty/\infty} \lesssim \sqrt{\frac{\tau \log d}{N_t}}$ *and* $\|\widehat{\Theta}_t - \Theta_t^*\|_F \lesssim \sqrt{\frac{\tau |\mathcal{S}_t| \log d}{N_t}}$.

Theorem 3 is a direct consequence of Theorem 1 and provides, to the best of our knowledge, the first non-asymptotic guarantee on the inference of sparsely-changing GMRFs with an arbitrary length of

time horizon T. In particular, it shows that the proposed optimization (3) guarantees small estimation error and zero mismatch error for sparsely-changing GMRFs, provided that N_t scales logarithmically with the dimension of the precision matrices. In the static setting ($T = 0$), the derived bound recovers the existing results on the sample complexity of learning static GMRFs [26, 34, 36].

4.2 Sparsely-and-smoothly-changing GMRFs

In many applications, such as financial markets and motion detection in video frames, the associated graphical model should be learned "on-the-go", as the data arrives with a continuously changing graphical model. Under such circumstances, one may have access to few (or even one) samples at each time.

Suppose that the precision matrices change smoothly over time. Such smooth changes can be modeled via a continuous function $\Theta(x) : [0, 1] \to \mathbb{R}^{d\times d}$ with uniformly bounded element-wise second derivatives $[\Theta(x)_{ij}]'' = \frac{d^2\Theta(t)_{ij}}{dx^2}$, such that $\Theta_t^* = \Theta(t/T)$ [16, 49]. If $\Theta(t) \succeq aI$ for every $t \in [0, 1]$ and some $a > 0$, then the covariance matrix $\Sigma(t) = \Theta(t)^{-1}$ is well-defined and smooth. Then, the problem of inferring the time-varying GMRF reduces to estimating a sequence of precision matrices $\{\Theta(0), \Theta(1/T), \ldots, \Theta(1)\}$ given the samples $X_t \sim \mathcal{N}(0, \Sigma(t/T))$. To alleviate the scarcity of samples, [16, 49] leverage the smoothness of the precision matrices, by taking the weighted average of the samples over time, where the weights are obtained from a nonparametric kernel. In particular, consider the weighted sample covariance matrix $\widehat{\Sigma}_t^w$:

$$\widehat{\Sigma}_t^w = \sum_{s=0}^{t} w(s,t)\Sigma_s, \text{ where } w(s,t) = \frac{1}{Th}K\left(\frac{s-t}{Th}\right) \tag{5}$$

and $K(\cdot)$ is a symmetric nonnegative kernel that satisfies a set of mild conditions which hold for most standard kernels, including the (truncated) Gaussian kernel. These conditions are delineated in the appendix. Next, we present the counterparts of Assumption 1 and Definition 1 for sparsely-and-smoothly-changing GMRFs.

Assumption 2 (Bounded norm)**.** *There exist constant numbers $\kappa_1 < \infty$, $\kappa_2 > 0$, and $\kappa_3 < \infty$ such that, for every $t \in [0, T]$,*

$$\|\Theta(t/T)\|_\infty \le \kappa_1, \quad \inf_{w:\|w\|_\infty=1} \|\Sigma(t/T)w\|_\infty \ge \kappa_2, \quad \|\Sigma(T)\|_{\infty/\infty} \le \kappa_3.$$

Definition 2 (Weak sparsity)**.** *Given $0 \le q \le 1$ and dimension d, define $s_c(q, d) \overset{def}{=} \max_{i,t\in[0,T]} \sum_{j=1}^{d} |[\Sigma(t/T)]_{ij}|^q$.*

We now present the analog of Theorem 3 for sparsely-and-smoothly-changing GMRFs.

Theorem 4. *Consider a sparsely-and-smoothly-changing GMRF with one sample per time, let $\zeta = \max\{\log_d(T + 1), 1\}$, and suppose that the sample covariance matrices are constructed according to* (5) *with $h \asymp T^{-1/3}$. Given an arbitrary $\tau > \zeta + 2$, let $T \gtrsim s_c(q, d)^{\frac{3}{1-q}}(\tau \log d)^{3/2}$. Then, with the approximate backward mapping $\widetilde{F}^*(\widehat{\Sigma}_t) = \mathtt{ST}_{\nu_t}(\widehat{\Sigma}_t^w)$ and parameters $\nu_t \asymp \frac{\sqrt{\tau \log d}}{T^{1/3}}$ and $\lambda_t \asymp \frac{\sqrt{\tau \log d}}{T^{1/3}}$, the estimates $\{\widehat{\Theta}_t\}_{t=0}^T$ obtained from* (3) *satisfy the following statements for all $T = 0, 1, \ldots, T$ with probability of at least $1 - d^{-\tau+\zeta+2}$:*

Sparsistency $\operatorname{supp}\left(\widehat{\Theta}_t\right) = \operatorname{supp}\left(\Theta_t^*\right)$ *and* $\operatorname{supp}\left(\widehat{\Theta}_t - \widehat{\Theta}_{t-1}\right) = \operatorname{supp}\left(\Theta_t^* - \Theta_{t-1}^*\right)$.

Estimation error $\|\widehat{\Theta}_t - \Theta_t^*\|_{\infty/\infty} \lesssim \sqrt{\frac{\tau \log d}{T^{2/3}}}$ *and* $\|\widehat{\Theta}_t - \Theta_t^*\|_F \lesssim \sqrt{\frac{\tau|\mathcal{S}_t| \log d}{T^{2/3}}}$.

Theorem 4 shows how the smoothness assumption on the true covariance matrix can be used to construct the backward mappings using the samples collected during the *entire* time horizon, thereby significantly reducing the sample complexity of learning time-varying GMRFs. In particular, leveraging the smoothness of the covariance matrix can reduce the minimum required number of samples from $\mathcal{O}(T \log d)$ to $\mathcal{O}((\log d)^{1.5})$. On the other hand, Theorem 3 does not impose any lower bound on T, and its estimation error decays faster in terms of the sample size.

5 Solution Method

In this section, we describe the proposed algorithm for solving (3). For the simplicity of notation, we define the lower bound and upper bound matrices l_t and u_t as $l_{t;ij} = [\widetilde{F}^*(\widehat{\Sigma}_t)]_{ij} - \lambda_t$ and $u_{t;ij} = [\widetilde{F}^*(\widehat{\mu}_t)]_{ij} + \lambda_t$, for every $1 \leq i, j \leq d$. The following fact plays a key role in our analysis.

Fact 1. *An optimal solution* $\left\{\widehat{\Theta}_t\right\}_{t=0}^T$ *of* (3) *satisfies for every* $1 \leq i < j \leq d$,

$$\left\{\widehat{\Theta}_{t;ij}\right\}_{t=0}^T \in \underset{\{\Theta_{t;ij}\}_{t=0}^T}{\arg\min} \quad (1-\gamma)\sum_{t=0}^T \mathbb{1}\{\Theta_{t;ij} \neq 0\} + \gamma \sum_{t=1}^T \mathbb{1}\{\Theta_{t;ij} - \Theta_{t-1;ij} \neq 0\} \tag{6a}$$

$$\text{s.t.} \;\; l_{t;ij} \leq \Theta_{t;ij} \leq u_{t;ij} \quad \forall 0 \leq t \leq T. \tag{6b}$$

Fact 1 implies that (3) decomposes into the smaller subproblems (6). Therefore, our main focus is devoted to solving each subproblem independently. To further simplify the notation, we drop the subscript ij from (6), whenever it is chosen arbitrarily, and use θ_t instead of $\Theta_{t;ij}$.

Let $\texttt{OPT}_{i\to j}(\gamma)$ denote the truncated problem from time i to j with the regularization coefficient γ:

$$f^*_{i\to j} = \min_{\{\theta_t\}_{t=i}^j} \quad (1-\gamma)\sum_{t=i}^j \mathbb{1}\{\theta_t \neq 0\} + \gamma \sum_{t=i+1}^j \mathbb{1}\{\theta_t - \theta_{t-1} \neq 0\} \tag{7a}$$

$$\text{subject to} \;\; l_t \leq \theta_t \leq u_t \quad \forall i \leq t \leq j. \tag{7b}$$

Let the objective function for a candidate solution θ be denoted as $f_{i\to j}(\theta)$; by convention, we let $f_{i\to j}(\theta) = 0$ whenever $j < i$. Moreover, the optimal objective value and the set of optimal solutions to $\texttt{OPT}_{i\to j}(\gamma)$ are respectively denoted as $f^*_{i\to j}$ and $\mathcal{X}^*_{i\to j}$. Similarly, $\widehat{\theta}_{i\to j} \in \mathcal{X}^*_{i\to j}$ is used to denote an optimal solution to $\texttt{OPT}_{i\to j}(\gamma)$. We omit the subscript $i \to j$ whenever $i = 0$ and $j = T$. The tth feasible interval is defined as $\Delta_t = [l_t, u_t]$. Accordingly, the notation $\Delta^{\cap}_{t\to s}$ refers to $\Delta^{\cap}_{t\to s} \stackrel{\text{def}}{=} \Delta_t \cap \Delta_{t+1} \cap \cdots \cap \Delta_s$.

5.1 Special case: $\gamma = 1$

As the first step, we consider the special case $\gamma = 1$, and provide an efficient algorithm (Algorithm 1) for solving $\texttt{OPT}_{0\to T}(1)$, where the sparsity is only promoted on the parameter differences (and not on the individual parameters). As will be shown later, Algorithm 1 will be used as a subroutine in our proposed algorithm for the general case $0 < \gamma < 1$. At a high level, Algorithm 1 recursively performs the following operations: at any given time τ, the algorithm looks into the future to find a nonempty interval that is feasible for the longest possible time δ. Then, it sets the subvector $\theta_{\tau:\delta}$ to an arbitrarily chosen element from this nonempty interval.

Algorithm 1 $\texttt{Greedy}(l, u, \tau, T)$

Output: Solution $\theta^{\texttt{Greedy}}_{\tau\to T}$, the objective value $f^{\texttt{Greedy}}_{\tau\to T}$ to $\texttt{OPT}_{\tau\to T}(1)$, and the index set Γ of maximal nonempty intervals
Find largest δ such that $\Delta^{\cap}_{\tau\to\delta} \neq 0$;
Set $\theta^{\texttt{Greedy}}_{\tau:\delta} = \eta$ for some $\eta \in \Delta^{\cap}_{\tau\to\delta}$;
if $\delta \leq T - 1$ **then**
    Execute $\texttt{Greedy}(l, u, \delta + 1, T)$;
end if
Set $f^{\texttt{Greedy}}_{\tau\to T} = \sum_{t=\tau+1}^T \mathbb{1}\{\theta^{\texttt{Greedy}}_t - \theta^{\texttt{Greedy}}_{t-1} \neq 0\}$;
Set $\Gamma \leftarrow \Gamma \cup \{\delta\}$;
Return $\{\theta^{\texttt{Greedy}}_t\}_{t=\tau}^T$, $f^{\texttt{Greedy}}_{\tau\to T}$, and Γ;

Proposition 1. *$\texttt{Greedy}(l, u, 0, T)$ returns an optimal solution $\{\theta^{\texttt{Greedy}}_t\}_{t=0}^T$ to $\texttt{OPT}_{0\to T}(1)$. Moreover, the truncated solution $\{\theta^{\texttt{Greedy}}_t\}_{t=0}^j$ is optimal for $\texttt{OPT}_{0\to j}(1)$.*

5.2 General case: $0 < \gamma < 1$

Now, we present our main algorithm for the general case $0 < \gamma < 1$.

Definition 3. *The set $\mathcal{Z}_{i\to j} = \{i, i+1, \dots, j\}$ is called a* **zero-feasible sequence** *if $l_k \le 0 \le u_k$ for every $k \in \mathcal{Z}_{i\to j}$. Moreover, the zero-feasible sequence $\mathcal{Z}_{i\to j}$ is called* **maximal** *if it is not strictly contained within another zero-feasible sequence.*

Let $\mathcal{Z}_{i_1\to j_1}, \mathcal{Z}_{i_2\to j_2}, \dots, \mathcal{Z}_{i_Z\to j_Z}$ be the set of all maximal zero-feasible sequences such that $0 \le i_1 \le j_1 < i_2 \le j_2 < \dots < i_Z \le j_Z \le T$, where Z is the number of maximal zero-feasible sequences. If $Z = 0$, i.e., there is no zero-feasible sequence, then it is easy to see that $\sum_{t=0}^{T} \mathbb{1}\{\theta_t \neq 0\} = T+1$ for every feasible solution, and hence, $\texttt{Greedy}(l, u, 0, T)$ leads to an optimal solution to (7). Another special case is when $i_1 = 0$ and $j_1 = T$, in which case the optimal solution is $\widehat{\theta} = 0$. Therefore, without loss of generality, suppose that $Z > 0$ and either $i_1 \neq 0$ or $j_1 \neq T$.

Our goal now is to obtain an optimal solution to (7) by solving a shortest path problem over a *weighted directed acyclic graph* (DAG) whose nodes correspond to the maximal zero-feasible sequences. In particular, consider a weighted DAG $\mathcal{G}$ with the vertex set $V = \{0, 1, \dots, Z, Z+1\}$, where the vertices k and l are connected via a directed arc (k, l) if $k < l$. Moreover, for every arc (k, l), the weight $W(k, l) = 0$ if $(k, l) = (0, 1), i_1 = 0$ or $(k, l) = (Z, Z+1), j_Z = T$, and

$$W(k,l) = (1-\gamma)(i_l - j_k - 1) + \gamma f^{\texttt{Greedy}}_{j_k+1\to i_l-1} + \gamma \mathbb{1}\{k \neq 0\} + \gamma \mathbb{1}\{l \neq Z+1\} \tag{8}$$

otherwise, where we define $j_0 = -1$ and $i_{Z+1} = T+1$.

Algorithm 2 Algorithm for solving (7)

1: **Output:** Optimal solution $\widehat{\theta}$ and the objective value f^* to $\texttt{OPT}_{0\to T}(\gamma)$
2: Find the maximal zero-feasible sequences;
3: Construct the DAG $\mathcal{G}$ with weights defined as (8);
4: Find the shortest path $p = (v_1, v_2, \dots, v_r)$ between the vertices 0 and $Z+1$ in $\mathcal{G}$;
5: Set $\widehat{\theta}_{j_{v_i}+1:i_{v_{i+1}}-1} = \theta^{\texttt{Greedy}}_{j_{v_i}+1\to i_{v_{i+1}}-1}$ and $\widehat{\theta}_{i_{v_l}:j_{v_{l+1}}} = 0$ for every $l = 1, 2, \dots, r$;
6: **Return** $\{\widehat{\theta}\}_{t=0}^{T}$ and f^*

Theorem 5. *The shortest path from* 0 *to* $Z+1$ *on* $\mathcal{G}$ *has value* f^*.

The above theorem implies that the optimal solution to (3) can be obtained via Algorithm 2.

Theorem 6. *Problem* (7) *can be solved in* $\mathcal{O}(ZT)$ *time and memory.*

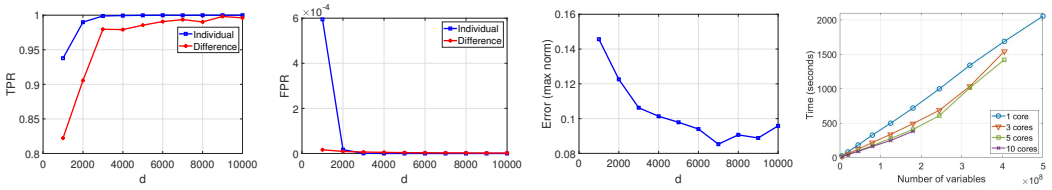
Since $Z = \mathcal{O}(T)$ and a solution to (3) requires solving $\mathcal{O}(d^2)$ instances of (7), we find the total complexity stated in Theorem 2. Note however that if $Z = \mathcal{O}(1)$, then the overall complexity reduces to $\mathcal{O}(Td^2)$, which is linear in the total number of variables. In the next section, we will show that the practical runtime of the proposed algorithm is near-linear with respect to the number of variables.

6 Numerical Analysis

In this section, we evaluate the performance of the proposed estimator in synthetically generated massive datasets, and a case study on the correlation network inference in stock markets. We refer the reader to the supplementary file for an extensive discussion of our simulations (e.g., the choice of parameters, robustness analysis, and a comparison with other state-of-the-art methods). In all of our simulations, the parameters ν_t and λ_t are chosen directly from the data samples, i.e., without prior knowledge of the true solution, via Bayesian Inference Criterion (BIC) [32, 14].

Case Study on Large Datasets We consider randomly generated instances of sparsely-changing GMRFs, where the true inverse covariance matrix is constructed as follows: at time $t = 0$, we set $\Theta_0 = I_d + \sum_{(i,j)\in\mathcal{S}} A^{(i,j)}$, where $A^{(i,j)}$ is a sparse positive semidefinite matrix with exactly two nonzero off-diagonal elements. For every $(i, j) \in \mathcal{S}$, we set $A^{(i,j)}_{ij} = A^{(i,j)}_{ji} = -0.4$ and $A^{(i,j)}_{ii} = A^{(i,j)}_{jj} = 0.4$. Clearly, $A^{(i,j)} \succeq 0$, and hence, $\Theta_0 \succ 0$. In the first experiment, we fix

$T = 10$ and change the values of d. The number of nonzero elements in the individual precision matrices and their differences are set to $3d$ and $0.04d$, respectively. We evaluate the performance of the proposed method in the high dimensional settings, where $N_t = d/2$ for every $t = 0, \ldots, T$. Moreover, define TPR and FPR for the individual parameters and their differences as the *true positive* and *false positive* values, normalized by the total number of nonzero and zero elements in the true precision matrices and their differences, respectively. Clearly, both TPR and FPR are between 0 and 1, with TPR $= 1$ and FPR $= 0$ corresponding to the perfect recovery of the sparsity patterns. Figure 7 depicts TPR, FPR, and the max-norm error of the estimated parameters, as well as the runtime of our algorithm for different values of d with and without parallelization. It can be seen that both TPR and FPR values improve with the dimension for the estimated parameters and their differences. Moreover, with a single core, the runtime of our algorithm scales almost linearly with d^2, which is in line with the result of Theorem 2. Using 5 cores, the runtime of our algorithm is improved by 40% on average. Using 10 cores deteriorates the performance due to the shared memory limitations. *Using our algorithm, we reliably infer instances of sparsely-changing GMRFs with more almost 500 million variables in less than one hour.*

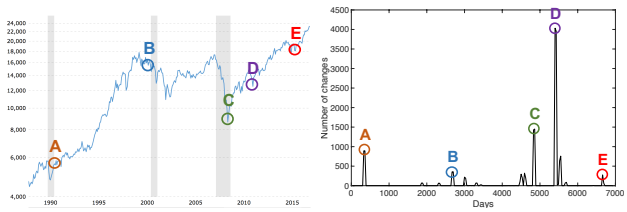


(a) (b) (c) (d)

Figure 3: (a) TPR of the proposed method. (b) FPR of the proposed method. (c) The max-norm of estimation error. (d) The runtime of the proposed algorithm with and without parallelization.

Case Study on Stock Market Finally, we illustrate the performance of our algorithm for the inference of stock correlation network. In particular, we consider an investor seeking to understand relationships between securities over time. Sparsity of the precision matrices guarantees interpretability, while the sparsely-changing structure is imposed to identify sharp changes in market conditions, and a need to rebalance the portfolio. We consider the daily changes for 214 securities from $1990/01/04$ to $2017/08/10$, with the total number of 6990 days ($d = 214$ and $T = 6990$). Due to the continuously changing nature of the stock correlation network, we will use the kernel averaging approach introduced in Subsection 4.1. Using the constructed sample covariance matrices, we estimate the sparsely-changing precision matrix $\Theta(t/T)$ at discrete times $t \in \{30, 60, 90, \ldots, 6990\}$.

A drastic change in the correlation network signals a *spike* in the stock market, which may reflect the market's response to unexpected events. Figure 11 shows the number of changes in the estimated network, for the choices of $\nu_0 = 3$, $\lambda_0 = 0.16$, and $\gamma = 0.9$, together with the historical chart of National Association of Securities Dealers Automated Quotations (NASDAQ) [1]. It can be seen that the major spikes in the estimated network can be attributed to the historical stock market crashes. For instance, the spikes A, B, and C respectively correspond to the "early 1990s recession", "dot-com bubble", and "global financial crisis"; see [5] for more details. Interestingly, the estimated network can also detect other historical (but less severe) downturns in 2011 (point D) and 2016 (point E). In the supplementary materials, we provide a more detailed analysis with different choices of parameters.



(a) (b)

Figure 4: (a) NASDAQ historical chart [2]. (b) The number of changes in the estimated network.

References

[1] About nasdaq. https://www.nasdaq.com/about, 2020.

[2] Nasdaq composite - 45 year historical chart. https://www.macrotrends.net/1320/nasdaq-historical-chart, 2020.

[3] Ravindra K Ahuja, Dorit S Hochbaum, and James B Orlin. A cut-based algorithm for the nonlinear dual of the minimum cost network flow problem. *Algorithmica*, 39(3):189–208, 2004.

[4] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Network flows. 1988.

[5] Robert Z Aliber and Charles P Kindleberger. *Manias, panics, and crashes: A history of financial crises*. Springer, 2017.

[6] Alper Atamtürk and Andrés Gómez. Strong formulations for quadratic optimization with m-matrices and indicator variables. *Mathematical Programming*, 170(1):141–176, 2018.

[7] Alper Atamtürk, Andrés Gómez, and Shaoning Han. Sparse and smooth signal estimation: Convexification of ℓ_0-formulations. *arXiv preprint arXiv:1811.02655*, 2018.

[8] Francis Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 175(1-2):419–459, 2019.

[9] Michele Benzi and Nader Razouk. Decay bounds and o (n) algorithms for approximating functions of sparse matrices. *Electron. Trans. Numer. Anal*, 28:16–39, 2007.

[10] Michele Benzi and Valeria Simoncini. Decay bounds for functions of hermitian matrices with banded or kronecker structure. *SIAM Journal on Matrix Analysis and Applications*, 36(3):1263–1282, 2015.

[11] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[12] Biao Cai, Gemeng Zhang, Aiying Zhang, Julia M Stephen, Tony W Wilson, Vince D Calhoun, and Yu-Ping Wang. Capturing dynamic connectivity from resting state fmri using time-varying graphical lasso. *IEEE Transactions on Biomedical Engineering*, 66(7):1852–1862, 2018.

[13] Stephen Demko, William F Moss, and Philip W Smith. Decay rates for inverses of band matrices. *Mathematics of computation*, 43(168):491–499, 1984.

[14] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *arXiv preprint arXiv:1011.6640*, 2010.

[15] Sinong Geng, Minhao Yan, Mladen Kolar, and Sanmi Koyejo. Partially linear additive gaussian graphical models. In *International Conference on Machine Learning*, pages 2180–2190. PMLR, 2019.

[16] Kristjan Greenewald, Seyoung Park, Shuheng Zhou, and Alexander Giessing. Time-dependent spatially varying graphical models, with application to brain fmri data analysis. In *Advances in Neural Information Processing Systems*, pages 5832–5840, 2017.

[17] David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 205–213, 2017.

[18] Dorit S Hochbaum. An efficient algorithm for image segmentation, markov random fields and related problems. *Journal of the ACM (JACM)*, 48(4):686–701, 2001.

[19] Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, and Eric Reiman. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[20] Sean Jewell and Daniela Witten. Exact spike train inference via ℓ_0 optimization. *arXiv preprint arXiv:1703.08644*, 2017.

[21] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[22] D Kershaw. Inequalities on the elements of the inverse of a certain tridiagonal matrix. *Mathematics of computation*, pages 155–158, 1970.

[23] Junghi Kim and Wei Pan. Highly adaptive tests for group differences in brain functional connectivity. *NeuroImage: Clinical*, 9:625–639, 2015.

[24] Mladen Kolar and Eric Xing. On time varying undirected graphs. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 407–415. JMLR Workshop and Conference Proceedings, 2011.

[25] Mladen Kolar and Eric P Xing. Estimating networks with jumps. *Electronic journal of statistics*, 6:2069, 2012.

[26] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[27] Batiste Le Bars, Pierre Humbert, Argyris Kalogeratos, and Nicolas Vayatis. Learning the piece-wise constant graph structure of a varying ising model. *International Conference on Machine Learning*, 2020.

[28] Song Liu, Kenji Fukumizu, and Taiji Suzuki. Learning sparse structural changes in high-dimensional markov networks. *Behaviormetrika*, 44(1):265–286, 2017.

[29] Xiao Liu and Jeff H Duyn. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 110(11):4392–4397, 2013.

[30] Alan Miller. *Subset Selection in Regression*. CRC Press, 2002.

[31] Manjari Narayan, Genevera I Allen, and Steffie Tomson. Two sample inference for populations of graphical models with applications to functional connectivity. *arXiv preprint arXiv:1502.03853*, 2015.

[32] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[33] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[34] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[35] Alessandro Rinaldo et al. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37:2922–2952, 2009.

[36] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[37] Sandipan Roy, Yves Atchadé, and George Michailidis. Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(4):1187, 2017.

[38] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[39] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

[40] Jerome Clair Shellenbarger. Estimation of covariance parameters for an adaptive kalman filter. 1966.

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

[42] Curtis R Vogel and Mary E Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17:227–238, 1996.

[43] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[44] Beilun Wang, Yanjun Qi, et al. Fast and scalable learning of sparse changes in high-dimensional gaussian graphical model structure. In *International Conference on Artificial Intelligence and Statistics*, pages 1691–1700. PMLR, 2018.

[45] Jialei Wang and Mladen Kolar. Inference for sparse conditional precision matrices. *arXiv preprint arXiv:1412.7638*, 2014.

[46] Yair Weiss and William T Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. In *Advances in Neural Information Processing Systems*, pages 673–679, 2000.

[47] Eunho Yang, Aurélie C Lozano, and Pradeep K Ravikumar. Elementary estimators for graphical models. In *Advances in neural information processing systems*, pages 2159–2167, 2014.

[48] Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

[49] Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the last paragraph of Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The detailed proofs are in the supplementary file.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] For synthetically generated experiments, we have carefully outlined the data generation process. For the real-world case study, we have included a reference to the used dataset. Finally, a full description of the hyper-parameters of the algorithm is provided.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]
(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [No]
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]