Graphical Nonconvex Optimization via an Adaptive Convex Relaxation

Qiang Sun [1] **Kean Ming Tan** [2] **Han Liu** [3] **Tong Zhang** [3]

Abstract

We consider the problem of learning high-dimensional Gaussian graphical models. The graphical lasso is one of the most popular methods for estimating Gaussian graphical models. However, it does not achieve the oracle rate of convergence. In this paper, we propose the graphical nonconvex optimization for optimal estimation in Gaussian graphical models, which is then approximated by a sequence of adaptive convex programs. Our proposal is computationally tractable and produces an estimator that achieves the oracle rate of convergence. The statistical error introduced by the sequential approximation is clearly demonstrated via a contraction property. The proposed methodology is then extended to modeling semiparametric graphical models. We show via numerical studies that the proposed estimator outperforms other popular methods for estimating Gaussian graphical models.

1. Introduction

We consider the problem of learning an undirected graph $G = (V, E)$, where $V = \{1, \ldots, d\}$ is a set of nodes that represents d random variables, and E is an edge set that describes the pairwise conditional dependence relationships among the d random variables. Gaussian graphical models have been widely used to represent pairwise conditional dependencies among a set of random variables. Let $\boldsymbol{X}$ be a d-dimensional random variables. Under the Gaussian assumption $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^*)$, the graph G is encoded by the sparse concentration matrix $\mathbf{\Theta}^* = (\mathbf{\Sigma}^*)^{-1}$, or the sparse inverse correlation matrix $\mathbf{\Psi}^* = (\mathbf{C}^*)^{-1}$. Here, $\mathbf{C}^*$ is the correlation matrix such that $\mathbf{\Sigma}^* = \mathbf{W}\mathbf{C}^*\mathbf{W}$ and $\mathbf{W}^2$ is a diagonal matrix with diagonal elements of $\mathbf{\Sigma}^*$. It is well known that the jth and kth variables are conditionally independent given all of the other variables if and only if the (j, k)-th element of $\mathbf{\Theta}^*$ (or $\mathbf{\Psi}^*$) is equal to zero. Thus, inferring the conditional dependencies structure of a Gaussian graphical model boils down to estimating a sparse inverse covariance (or correlation) matrix.

A number of methods have been proposed to estimate the sparse concentration matrix under the Gaussian assumption. For example, Meinshausen & Bühlmann (2006) proposed a neighborhood selection approach for estimating Gaussian graphical models by solving a collection of sparse linear regression problems using the lasso penalty. In addition, Yuan (2010) and Cai et al. (2011) proposed the graphical Dantzig and CLIME, both of which can be solved efficiently. From a different perspective, Yuan & Lin (2007) and Friedman et al. (2008) proposed the graphical lasso, a penalized likelihood based approach, to estimate the concentration matrix $\mathbf{\Theta}^*$ directly. Various extensions of the graphical lasso were proposed and the theoretical properties were also studied (among others, Banerjee et al., 2008; Rothman et al., 2008; Ravikumar et al., 2011). The Gaussian graphical models literature is vast and we refer the reader to Cai et al. (2016a) and Drton & Maathuis (2016) for a comprehensive review.

Despite the popularity of the graphical lasso on modeling sparse Gaussian graphical models, it does not achieve the oracle rate of convergence. More specifically, it is believed that the optimal rate of convergence in spectral norm for the graphical lasso is at the order of $\sqrt{s \log d / n}$ (Rothman et al., 2008). Here, n is the sample size, d is the number of nodes, and s is the number of edges in the true graph. In fact, the graphical lasso and all of the aforementioned methods are based on the lasso penalty and it is generally believed that convex penalties usually introduce non-negligible estimation bias. For example, in the linear regression setting, Fan & Li (2001); Zhang (2010a;b); Fan et al. (2018) have shown that the nonconvex penalized regression is able to eliminate the estimation bias and attain a more refined statistical rate of convergence.

Based on these insights, we propose the following penalized maximum likelihood estimation with a general nonconvex

[1]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada [2]School of Statistics, University of Minnesota, Minneapolis, MN, USA [3]Tencent AI Lab, Tencent Technology, Shenzhen, China. Correspondence to: Qiang Sun <qsun@utstat.toronto.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

penalty:

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathcal{S}_+^d}{\operatorname{argmin}} \left\{ \langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}} \rangle - \log\det(\boldsymbol{\Theta}) + \sum_{i \neq j} p_\lambda(\Theta_{ij}) \right\}, \quad (1.1)$$

where $\mathcal{S}_+^d = \{\mathbf{A} \in \mathbb{R}^{d \times d} : \mathbf{A} = \mathbf{A}^{\mathrm{T}}, \mathbf{A} \succ 0\}$ is the symmetric definite cone formed by all $d \times d$ symmetric positive definite matrices, $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix, and $p_\lambda(\cdot)$ is a nonconvex penalty. Here, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ denotes the trace of $\mathbf{A}^{\mathrm{T}}\mathbf{B}$. However, from the computational perspective, minimizing a penalized loss function with nonconvex penalty is a challenging problem due to its intrinsic nonconvex structure. For example, Ge et al. (2011) have shown that solving (1.1) with the ℓ_p penalty is strongly NP-hard, when $0 \leq p < 1$. In other words, there does not exist a fully polynomial-time approximation scheme for problem (1.1) unless more structures are assumed.

Recently, Loh & Wainwright (2015) proposed an algorithm to obtain a good local optimum for regression problems similar to (1.1), under an additional convex constraint that depends on the unknown true parameters. Loh & Wainwright (2015) have established estimation error under various vector norm such as the ℓ_2 and ℓ_∞. However, Loh & Wainwright (2015) failed to provide a faster rate of convergence statistically due to not taking signal strength into account. Computationally, our algorithm is different from the path-following algorithm in Wang et al. (2014), which must start from the largest regularization parameter. Our algorithm directly starts form the target regularization parameter, which is in the order of $\sqrt{\log d/n}$. This could help in some cases. For example, if we have some prior knowledge about the range of λ, then we do not need to start from the largest regularization parameter. In the context of Gaussian graphical models, the rate of convergence of $\sqrt{s \log d/n}$ under the spectral norm has been obtained in the existing literature. In this paper, we further improve the rate to $\sqrt{s/n}$. To the best of our knowledge, we are the first in the literature to obtain this sharp rate under the operator norm.

In this paper, instead of directly solving the nonconvex problem (1.1), we propose to approximate it by a sequence of adaptive convex programs. Even though the proposed method involves solving a sequence of convex programs, we show that the proposed estimator for estimating the sparse concentration matrix achieves the oracle rate of convergence of $\sqrt{s/n}$, as if the locations of the nonzeros in the sparse concentration matrix were known a priori. This is achieved by a contraction property. Roughly speaking, each convex program gradually contracts the initial estimator to the region of oracle rate of convergence even when a bad initial estimator is used in the first place:

$$\|\widehat{\boldsymbol{\Psi}}^{(\ell)} - \boldsymbol{\Psi}^*\|_{\mathrm{F}} \leq \underbrace{C\sqrt{\frac{s}{n}}}_{\text{Oracle Rate}} + \underbrace{\frac{1}{2}\|\widehat{\boldsymbol{\Psi}}^{(\ell-1)} - \boldsymbol{\Psi}^*\|_{\mathrm{F}}}_{\text{Contraction Effect}},$$

where $\widehat{\boldsymbol{\Psi}}^{(\ell)}$ is the inverse correlation matrix estimator after the ℓ-th convex approximation, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, C is a positive constant, and $\sqrt{s/n}$ is referred to as the oracle rate. Each iteration of the proposed method helps improve the accuracy only when $\|\widehat{\boldsymbol{\Psi}}^{(\ell-1)} - \boldsymbol{\Psi}^*\|_{\mathrm{F}}$ dominates the statistical error. The error caused by each iteration is clearly demonstrated via the proven contraction property. Suprisingly, we only need to solve about $\log\log d$ convex programs to achieve the oracle rate. By rescaling the inverse correlation matrix using the estimated marginal variances, we obtain an estimator of the concentration matrix with spectral norm convergence rate in the order of $\sqrt{\log d/n} \vee \sqrt{s/n}$, where $a \vee b = \max\{a, b\}$. By exploiting the sparsity pattern matrix of $\boldsymbol{\Theta}^*$, we further sharpen the rate of convergence to $\sqrt{s/n}$ under the spectral norm.

The rest of this paper proceeds as follows. Our proposed method and its implementation are detailed in Section 2. Section 3 is devoted to theoretical studies. We show that the proposed methodology can be extended to the semiparametric graphical models in Section 4. Numerical experiments are provided to support the proposed method in Section 5. We conclude the paper in Section 6. Proofs and technical details are in the supplementary material.

Notation: We summarize the notation that will be used regularly throughout the paper. Given a vector $\mathbf{u} = (u_1, u_2, \ldots, u_d)^{\mathrm{T}} \in \mathbb{R}^d$, we define the ℓ_q-norm of $\mathbf{u}$ by $\|\mathbf{u}\|_q = (\sum_{j=1}^d |u_j|^q)^{1/q}$, where $q \in [1, \infty)$. For a set $\mathcal{A}$, let $|\mathcal{A}|$ denote its cardinality. For a matrix $\mathbf{A} = (a_{i,j}) \in \mathbb{R}^{d \times d}$, we use $\mathbf{A} \succ 0$ to indicate that $\mathbf{A}$ is positive definite. For $q \geq 1$, we use $\|\mathbf{A}\|_q = \max_{\mathbf{u}} \|\mathbf{A}\mathbf{u}\|_q / \|\mathbf{u}\|_q$ to denote the operator norm of $\mathbf{A}$. For index sets $\mathcal{I}, \mathcal{J} \subseteq \{1, \ldots, d\}$, we define $\mathbf{A}_{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{d \times d}$ to be the matrix whose (i,j)-th entry is equal to $a_{i,j}$ if $i \in \mathcal{I}$ and $j \in \mathcal{J}$, and zero otherwise. We use $\mathbf{A} \odot \mathbf{B} = (a_{ij} b_{ij})$ to denote the Hadamard product of two matrices $\mathbf{A}$ and $\mathbf{B}$. Let $\mathrm{diag}(\mathbf{A})$ denote the diagonal matrix consisting diagonal elements of $\mathbf{A}$. We use $\mathrm{sign}(x)$ to denote the sign of x: $\mathrm{sign}(x) = x/|x|$ if $x \neq 0$ and $\mathrm{sign}(x) = 0$ otherwise. For two scalars f_n and g_n, we use $f_n \gtrsim g_n$ to denote the case that $f_n \geq c g_n$, and $f_n \lesssim g_n$ if $f_n \leq C g_n$, for two positive constants c and C. We say $f_n \asymp g_n$, if $f_n \gtrsim g_n$ and $f_n \lesssim g_n$. $\mathcal{O}_{\mathbb{P}}(\cdot)$ is used to denote bounded in probability. We use c and C to denote constants that may vary from line to line.

2. Graphical Nonconvex Optimization

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^{\mathrm{T}}$ be a mean zero d-dimensional Gaussian random vector. Then its density can be parameterized by the concentration matrix $\boldsymbol{\Theta}^*$ or the inverse correlation matrix $\boldsymbol{\Psi}^*$. The family of Gaussian distributions respects the edge structure of a graph $G = (V, E)$ in the sense that $\Psi^*_{ij} = 0$ if and only if $(i, j) \notin E$. This family is known as the Gauss-Markov random field with respect to the graph G.

Given n independent and identically distributed observations $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$ of a mean zero d-dimensional random vector $\boldsymbol{X} \in \mathbb{R}^d$, we are interested in estimating the inverse correlation matrix $\boldsymbol{\Psi}^*$ and concentration matrix $\boldsymbol{\Theta}^*$. Let $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}^{(i)}(\boldsymbol{X}^{(i)})^{\mathrm{T}}$ be the sample covariance matrix and let $\widehat{\mathbf{C}} = \widehat{\mathbf{W}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{W}}^{-1}$, where $\widehat{\mathbf{W}}^2 = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}})$. To estimate $\boldsymbol{\Psi}^*$, we propose to adaptively solve the following sequence of convex programs

$$\widehat{\boldsymbol{\Psi}}^{(\ell)} = \mathop{\mathrm{argmin}}_{\boldsymbol{\Psi}\in\mathcal{S}_+^d} \Big\{ \big\langle \boldsymbol{\Psi}, \widehat{\mathbf{C}} \big\rangle - \log\det(\boldsymbol{\Psi}) + \|\boldsymbol{\lambda}^{(\ell-1)} \odot \boldsymbol{\Psi}\|_{1,\mathrm{off}} \Big\}, \text{ for } \ell = 1, \ldots, T, \quad (2.1)$$

where $\|\boldsymbol{\Psi}\|_{1,\mathrm{off}} = \sum_{i\neq j}|\Psi_{ij}|$, $\boldsymbol{\lambda}^{(\ell-1)} = \lambda \cdot \mathrm{w}\big(\widehat{\Psi}^{(\ell-1)}_{ij}\big)$ is a $d \times d$ adaptive regularization matrix for a given tuning parameter λ and a weight function $\mathrm{w}(\cdot)$, and T indicates the total number of convex programs needed. The weight function $\mathrm{w}(\cdot)$ can be taken to be $\mathrm{w}(t) = p'_\lambda(t)/\lambda$, where $p_\lambda(t)$ is a folded concave penalty such as the SCAD or the MCP proposed by Fan & Li (2001) and Zhang (2010a), respectively.

To obtain an estimator for the concentration matrix $\boldsymbol{\Theta}^*$, we rescale $\widehat{\boldsymbol{\Psi}}^{(T)}$ back to $\widetilde{\boldsymbol{\Theta}}^{(T)} = \widehat{\mathbf{W}}^{-1}\widehat{\boldsymbol{\Psi}}^{(T)}\widehat{\mathbf{W}}^{-1}$ after the T-th convex program. This rescaling helps improve the rate of convergence for $\widetilde{\boldsymbol{\Theta}}^{(T)}$ significantly by eliminating the effect introduced through the unpenalized diagonal elements. The detailed routine is summarized in Algorithm 1.

The computational complexity of Step 2 in Algorithm 1 is $O(d^3)$: this is the complexity of the algorithm for solving the graphical lasso problem. We will show in the latter section that the number of iterations of Algorithm 1 can be chosen to be $T \approx \log\log d$ based on our theoretical analysis, yielding a computational complexity of $O(\log[\log(d)]d^3)$. Algorithm 1 can be implemented using existing `R` packages such as `glasso`. We note that our algorithm is an adaptive version of the SPICE algorithm in Rothman et al. (2008).

3. Theoretical Results

In this section, we study the theoretical properties of the proposed estimator. We start with some assumptions needed for the theoretical analysis.

Algorithm 1 A sequential convex approximation for the graphical nonconvex optimization.

Input: Sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, regularization parameter λ.

Step 1: Obtain sample correlation matrix $\widehat{\mathbf{C}}$ by $\widehat{\mathbf{C}} = \widehat{\mathbf{W}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{W}}^{-1}$, where $\widehat{\mathbf{W}}^2$ is a diagonal matrix with diagonal elements of $\widehat{\boldsymbol{\Sigma}}$.

Step 2: Solve a sequence of graphical lasso problems adaptively

$$\widehat{\boldsymbol{\Psi}}^{(\ell)} = \mathop{\mathrm{argmin}}_{\boldsymbol{\Psi}\in\mathcal{S}_+^d} \Big\{ \langle \boldsymbol{\Psi}, \widehat{\mathbf{C}} \rangle - \log\det(\boldsymbol{\Psi}) + \|\boldsymbol{\lambda}^{(\ell-1)} \odot \boldsymbol{\Psi}\|_{1,\mathrm{off}} \Big\},$$

$$\text{and } \boldsymbol{\lambda}^{(\ell)} = \lambda \cdot \mathrm{w}(\widehat{\Psi}^{(\ell)}_{ij}), \text{ for } \ell = 1, \ldots, T.$$

Step 3: Obtain an estimator of $\boldsymbol{\Theta}^*$ by $\widetilde{\boldsymbol{\Theta}}^{(T)} = \widehat{\mathbf{W}}^{-1}\widehat{\boldsymbol{\Psi}}^{(T)}\widehat{\mathbf{W}}^{-1}$.

3.1. Assumptions

Let $S = \big\{(i,j) : \Theta^*_{ij} \neq 0, i \neq j\big\}$ be the support set of the off-diagonal elements in $\boldsymbol{\Theta}^*$. Thus, S is also the support set of the off-diagonal elements in $\boldsymbol{\Psi}^*$. The first assumption we need concerns the structure of the true concentration and covariance matrices.

Assumption 3.1 (Structural Assumption)**.** We assume that $|S| \leq s, \|\boldsymbol{\Sigma}^*\|_\infty \leq M < \infty$, $0 < \varepsilon_1 \leq \sigma_{\min} \leq \sigma_{\max} \leq 1/\varepsilon_1 < \infty$, $0 < \varepsilon_2 \leq \lambda_{\min}(\boldsymbol{\Theta}^*) \leq \lambda_{\max}(\boldsymbol{\Theta}^*) \leq 1/\varepsilon_2 < \infty$. Here, $\sigma^2_{\max} = \max_j \Sigma^*_{jj}$ and $\sigma^2_{\min} = \min_j \Sigma^*_{jj}$, where $\boldsymbol{\Sigma}^* = \big(\Sigma^*_{ij}\big)$.

Assumption 3.1 is standard in the existing literature for Gaussian graphical models (see, for instance, Meinshausen & Bühlmann, 2006; Yuan, 2010; Cai et al., 2016b; Yuan & Lin, 2007; Ravikumar et al., 2011). We need $\sigma_{\min}$ and $\sigma_{\max}$ to be bounded from above and below to guarantee reasonable performance of the concentration matrix estimator (Rothman et al., 2008). Throughout this section, we treat $M, \varepsilon_1, \varepsilon_2$ as constants to simplify the presentation.

The second assumption concerns the weight functions, which are used to adaptively update the regularizers in Step 2 of Algorithm 1. Define the following class of weight functions:

$$\mathcal{W} = \Big\{ \mathrm{w}(t) : \mathrm{w}(t) \text{ is nonincreasing}, \; 0 \leq \mathrm{w}(t) \leq 1 \text{ if } t \geq 0, \mathrm{w}(t) = 1 \text{ if } t \leq 0 \Big\}. \quad (3.1)$$

Assumption 3.2 (Weight Function)**.** There exists an α such that the weight function $\mathrm{w}(\cdot) \in \mathcal{W}$ satisfies $\mathrm{w}(\alpha\lambda) = 0$ and $\mathrm{w}(u) \geq 1/2$, where $u = c\lambda$ for some constant c.

The above assumption on the weight functions can be easily satisfied. For example, it can be satisfied by simply taking $\mathrm{w}(t) = p'_\lambda(t)/\lambda$, where $p_\lambda(t)$ is a folded concave penalty such as the SCAD or the MCP (Fan & Li, 2001; Zhang, 2010a). Next, we impose an assumption on the magnitude of the nonzero off-diagonal entries in the inverse correlation matrix $\mathbf{\Psi}^*$.

Assumption 3.3 (Minimal Signal Strength)**.** The minimal signal satisfies $\min_{(i,j)\in S} \Psi^*_{ij} \geq (\alpha + c)\lambda \gtrsim \lambda$, where $c > 0$ is the same constant that appears in Assumption 3.2.

Assumption 3.3 is a mild condition. In the sub-Gaussian design case, λ can be taken to be the order of $\sqrt{\log d/n}$, which diminishes quickly as n increases. It is an analogue to the minimal signal strength assumption frequently assumed in nonconvex penalized regression problems (Fan & Li, 2001; Zhang, 2010a). Taking the signal strength into account, we can then obtain the oracle rate of convergence.

3.2. Main Theory

We now present several main theorems concerning the rates of convergence of the proposed estimator for the sparse inverse correlation and the concentration matrices, respectively. The following theorem concerns the rate of convergence for the one-step estimator $\widehat{\mathbf{\Psi}}^{(1)}$ obtained from Algorithm 1 when $\ell = 1$.

Proposition 3.4 (One-step Estimator)**.** Let $\lambda \asymp \sqrt{\log d/n}$. Under Assumption 3.1, we have

$$\big\|\widehat{\mathbf{\Psi}}^{(1)} - \mathbf{\Psi}^*\big\|_{\mathrm{F}} \lesssim \sqrt{\frac{s\log d}{n}}$$

with probability at least $1 - 8/d$,

Proof of Proposition 3.4. We collect the proof of Proposition 3.4 in Appendix A in the supplementary material. □

The above proposition indicates that the statistical error under the Frobenius norm for the one-step estimator is at the order of $\sqrt{s\log d/n}$, which is believed to be unimprovable when one-step convex regularization is used (Rothman et al., 2008; Ravikumar et al., 2011). However, when a sequence of convex programs is used as in our proposal, the rate of convergence can be improved significantly. This is demonstrated in the following theorem.

Theorem 3.5 (Contraction Property)**.** Suppose that $n \gtrsim s\log d$ and select λ such that $\lambda \asymp \sqrt{\log d/n}$. Under Assumptions 3.1, 3.2 and 3.3, with probability at least $1 - 8/d$, $\widehat{\mathbf{\Psi}}^{(\ell)}$ satisfies the following contraction property:

$$\big\|\widehat{\mathbf{\Psi}}^{(\ell)} - \mathbf{\Psi}^*\big\|_{\mathrm{F}} \leq \underbrace{8\|\mathbf{\Psi}^*\|_2^2\|\nabla\mathcal{L}(\mathbf{\Psi}^*)_S\|_{\mathrm{F}}}_{\text{Oracle Rate}} + \underbrace{\frac{1}{2}\big\|\widehat{\mathbf{\Psi}}^{(\ell-1)} - \mathbf{\Psi}^*\big\|_{\mathrm{F}}}_{\text{Contraction}},$$

for $1 \leq \ell \leq T$. Moreover, if $T \gtrsim \log(\lambda\sqrt{n}) \gtrsim \log\log d$, we have

$$\big\|\widehat{\mathbf{\Psi}}^{(T)} - \mathbf{\Psi}^*\big\|_{\mathrm{F}} = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{s}{n}}\bigg).$$

Proof of Theorem 3.5. The proof is collected in Appendix A in the supplementary material. □

Theorem 3.5 establishes a contraction property: each convex approximation contracts the initial estimator towards the true sparse inverse correlation matrix until it reaches the oracle rate of convergence, $\sqrt{s/n}$. To achieve the oracle rate, we need to solve no more than approximately $\log\log d$ convex programs. Note that $\log\log d$ grows very slowly as d increases and thus, in practice, we only need to solve a few convex programs to get a better estimator than existing method such as the graphical lasso. The rate of convergence $\sqrt{s/n}$ is better than the existing literature on likelihood-based methods for estimating sparse inverse correlation matrices (Rothman et al., 2008; Lam & Fan, 2009; Ravikumar et al., 2011). By rescaling, we obtain a concentration matrix estimator with a faster rate of convergence.

Theorem 3.6 (Faster Rate in Spectral Norm)**.** Under the same conditions as in Theorem 3.5, we have

$$\big\|\widetilde{\mathbf{\Theta}}^{(T)} - \mathbf{\Theta}^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{s}{n}} \vee \sqrt{\frac{\log d}{n}}\bigg).$$

Proof of Theorem 3.6. The proof is deferred to Appendix A in the supplementary material. □

The theorem above provides the optimal statistical rate for estimating sparse concentration matrices using likelihood based methods (Rothman et al., 2008; Lam & Fan, 2009; Ravikumar et al., 2011). The extra $\log d$ term is a consequence of estimating the marginal variances.

Definition 3.7 (Sparsity Pattern Matrix)**.** For a matrix $\mathbf{A} = \big(a_{ij}\big)$, we say $\mathbf{A}_{\mathrm{sp}} = \big(a^{\mathrm{sp}}_{ij}\big)$ is the corresponding sparsity pattern matrix if $a^{\mathrm{sp}}_{ij} = 1$ when $a_{ij} \neq 0$; and $a^{\mathrm{sp}}_{ij} = 0$, otherwise.

Let $\mathbf{M}^*$ be the sparsity pattern matrix of $\mathbf{\Psi}^*$ or $\mathbf{\Theta}^*$. Our next theorem provides an improved rate of convergence.

Theorem 3.8 (Improved Convergence Rate)**.** Suppose that $n \gtrsim (s + s_{\max}^2)\log d$ and take λ such that $\lambda \asymp \sqrt{\log d/n}$. Let $T \gtrsim \log s$. Under Assumptions 3.1, 3.2 and 3.3, we have

$$\big\|\widehat{\mathbf{\Psi}}^{(T)} - \mathbf{\Psi}^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\bigg(\|\mathbf{M}^*\|_2\sqrt{\frac{1}{n}}\bigg),$$
$$\big\|\widetilde{\mathbf{\Theta}}^{(T)} - \mathbf{\Theta}^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\bigg(\|\mathbf{M}^*\|_2\sqrt{\frac{1}{n}} \vee \sqrt{\frac{\log d}{n}}\bigg).$$

Proof of Theorem 3.8. The proof is deferred to Appendix B in the supplementary material. □

Theorem 3.8 suggests that the rates of convergence can be bounded using the spectral norm of the sparsity pattern matrix $\mathbf{M}^*$, which can be much sharper than those provided in Theorems 3.5 and 3.6. To demonstrate this observation, we consider a sequence of chain graphs specified by the following sparsity pattern matrices:

$$\mathbf{M}_k^c = \begin{bmatrix} \mathbf{A}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k-1} \end{bmatrix}, \quad \text{for} \quad k = 4, \ldots, 50,$$

where $\mathbf{A}_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is such that the (i,j)-th entry $A_{k,ij} = 1$ if $|i-j| \le 1$, and $A_{k,ij} = 0$ otherwise. $\mathbf{I}_{d-k-1} \in \mathbb{R}^{(d-k-1)\times(d-k-1)}$ is the identity matrix. Let s_k be the total sparsity of $\mathbf{M}_k^c$, that is $s_k = 2k$. We plot the ratio of the two rates of convergence for estimating $\mathbf{\Psi}^*$ in Theorems 3.5 and 3.8, $\|\mathbf{M}_k^c\|_2^2/s_k$, versus s_k in Figure 1. From Figure 1, we can see that the ratio goes to 0 as the total sparsity increases. This demonstrates that the convergence rate in Theorem 3.8 is indeed much sharper than that in Theorem 3.5, as least for the chain graphs constructed above. We also observe similar but less significant improvement for star-shape graphs. In Figure 2, we give an geometric illustration of the star and chain graphs.

4. Extension to Semiparametric Graphical Models

In this section, we extend the proposed method to modeling semiparametric graphical models. We focus on the nonparanormal family proposed by Liu et al. (2012), which is a nonparametric extension of the normal family. More specifically, we replace the random variable $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ by the transformed variable $f(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))^{\mathrm{T}}$, and assume that $f(\boldsymbol{X})$ follows a multivariate Gaussian distribution.

Definition 4.1 (Nonparanormal)**.** Let $f = \{f_1, \ldots, f_d\}^{\mathrm{T}}$ be a set of monotone univariate functions and let $\mathbf{\Sigma}^{\mathrm{npn}} \in \mathbb{R}^{d\times d}$ be a positive-definite correlation matrix with $\mathrm{diag}(\mathbf{\Sigma}^{\mathrm{npn}}) = \mathbf{1}$. A d-dimensional random variable $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ has a nonparanormal distribution $\boldsymbol{X} \sim \mathrm{NPN}_d(f, \mathbf{\Sigma}^{\mathrm{npn}})$ if $f(\boldsymbol{X}) \equiv (f(X_1), \ldots, f_d(X_d))^{\mathrm{T}} \sim N_d(\mathbf{0}, \mathbf{\Sigma}^{\mathrm{npn}})$.

We aim to recover the precision matrix $\mathbf{\Theta}^{\mathrm{npn}} = (\mathbf{\Sigma}^{\mathrm{npn}})^{-1}$. The main idea behind this procedure is to exploit Kendall's tau statistics to directly estimate $\mathbf{\Theta}^{\mathrm{npn}}$, without explicitly calculating the marginal transformation functions $\{f_j\}_{j=1}^d$. We consider the following Kendall's tau statistic:

$$\widehat{\tau}_{jk} = \frac{2\sum_{1\le i<i'\le n}\mathrm{sign}\big((X_j^{(i)} - X_j^{(i')})(X_k^{(i)} - X_k^{(i')})\big)}{n(n-1)}.$$

The Kendall's tau statistic $\widehat{\tau}_{jk}$ represent the nonparametric correlations between the empirical realizations of random variables X_j and X_k and is invariant to monotone transformations. Let $\widetilde{X}_j$ and $\widetilde{X}_k$ be two independent copies of X_j and X_k. The population version of Kendall's tau is given by $\tau_{jk} \equiv \mathrm{Corr}\big(\mathrm{sign}(X_j - \widetilde{X}_j), \mathrm{sign}(X_k - \widetilde{X}_k)\big)$. We need the following lemma which is taken from (Liu et al., 2012). It connects the Kendall's tau statistics to the underlying Pearson correlation coefficient $\mathbf{\Sigma}^{\mathrm{npn}}$.

Lemma 4.2. Assuming $\boldsymbol{X} \sim \mathrm{NPN}_d(f, \mathbf{\Sigma})$, we have $\Sigma_{jk}^0 = \sin\big(\tau_{jk}\cdot\pi/2\big)$.

Motivated by this Lemma, we define the following estimators $\widehat{\boldsymbol{S}} = [\widehat{S}_{jk}]$ for the unknown correlation matrix $\mathbf{\Sigma}^{\mathrm{npn}}$:

$$\widehat{S}_{jk}^{\tau} = \begin{cases} \sin\big(\widehat{\tau}_{jk}\cdot\pi/2\big), & j \ne k, \\ 1, & j = k. \end{cases}$$

Now we are ready to prove the optimal spectral norm rate for the Gaussian copula graphical model. The results are provided in the following theorem.

Theorem 4.3. Assume that $n \gtrsim s\log d$ and let $\lambda \asymp \sqrt{\log d/n}$. Under Assumptions 3.1, 3.2 and 3.3, $\widehat{\mathbf{\Theta}}^{(\ell)}$ satisfies the following contraction property:

$$\big\|\widehat{\mathbf{\Theta}}^{(\ell)} - \mathbf{\Theta}^*\big\|_{\mathrm{F}} \le \underbrace{4\|\mathbf{\Theta}^*\|_2^2\|\nabla\mathcal{L}(\mathbf{\Theta}^*)_S\|_{\mathrm{F}}}_{\text{Optimal Rate}} + \underbrace{\frac{1}{2}\big\|\widehat{\mathbf{\Theta}}^{(\ell-1)} - \mathbf{\Theta}^*\big\|_{\mathrm{F}}}_{\text{Contraction}}, \quad 1 \le \ell \le T,$$

with probability at least $1 - 8/d$. If $T \gtrsim \log(\lambda\sqrt{n}) \gtrsim \log\log d$, we have

$$\big\|\widehat{\mathbf{\Theta}}^{(T)} - \mathbf{\Theta}^*\big\|_{\mathrm{F}} = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{s}{n}}\bigg).$$

Proof of Theorem 4.3. The proof is deferred to Appendix C in the supplementary material. □

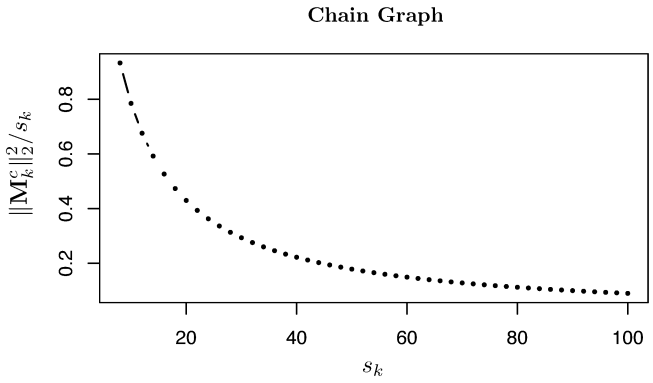

Figure 1. Convergence rates using sparsity pattern matrix $\mathbf{M}_k^c$ and total sparsity s_k.

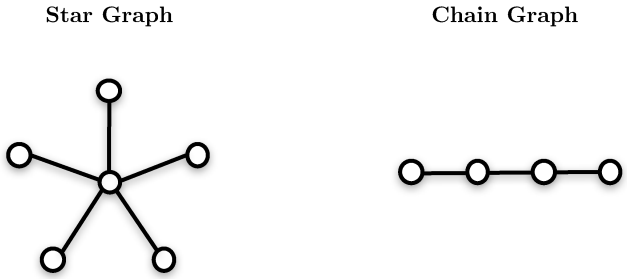

Figure 2. An illustration of the star and chain graphs.

5. Numerical Experiments

We compare our proposal to the graphical lasso (glasso) (Friedman et al., 2008) and neighborhood selection (NS) (Meinshausen & Bühlmann, 2006). Each of these approaches learns a Gaussian graphical model *via* an ℓ_1 penalty on each edge. To evaluate the performance across different methods, we define the true positive rate as the proportion of correctly identified edges in the graph, and the false positive rate as the proportion of incorrectly identified edges in the graph. In addition, we calculate the difference between the estimated and true concentration matrix under the Frobenius norm. We do not compute this quantity for the NS approach since they do not estimate the concentration matrix directly.

For our proposal, we consider $T = 4$ iterations with the SCAD penalty proposed by Fan & Li (2001) that takes the following form:

$$p'_\lambda(t) = \begin{cases} \lambda & \text{if } |t| \leq \lambda, \\ \frac{\gamma\lambda - |t|}{\gamma - 1} & \text{if } \lambda < |t| < \gamma\lambda, \\ 0 & \text{otherwise,} \end{cases}$$

where $\gamma > 2$. In all of our simulation studies, we pick $\gamma = 2.1$. Each of the methods involves a sparsity tuning parameter: we applied a fine grid of tuning parameter values to obtain the curves shown in Figure 3.

We consider cases with $n = \{150, 200\}$ and $d = 150$ with two set-ups for a $p \times p$ sparsity pattern matrix $\mathbf{A}$: (i) random graph with 2.5% elements of $\mathbf{A}$ set to 1; (ii) band graph with $A_{i,i+1} = A_{i+1,i} = 1$ for $1 \leq i \leq d-1$. We then use the sparsity pattern matrix $\mathbf{A}$ to create a matrix $\mathbf{E}$, as

$$E_{ij} = \begin{cases} 0 & \text{if } A_{ij} = 0 \\ 0.4 & \text{otherwise,} \end{cases}$$

and set $\mathbf{E} = \frac{1}{2}(\mathbf{E} + \mathbf{E}^T)$. Given the matrix $\mathbf{E}$, we set $\mathbf{\Theta}^{-1}$ equal to $\mathbf{E} + (0.1 - e_{\min})\mathbf{I}$, where $e_{\min}$ is the smallest eigenvalue of $\mathbf{E}$. We then standardize the matrix $\mathbf{\Theta}^{-1}$ so that the diagonals are equal to one. Finally, we generate the data according to $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)} \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$. We present the results averaged over 100 data sets for each of the two simulation settings with $n = \{150, 200\}$ and $d = 150$ in Figure 3.

From Row I of Figure 3, we see that our proposal is very competitive relative to the existing proposals for estimating Gaussian graphical models in terms of true and false positive rates across all simulation settings. Row II of Figure 3 contains the difference between the estimated and the true inverse covariance matrices under the Frobenius norm as a function of the false positive rate. For random graph

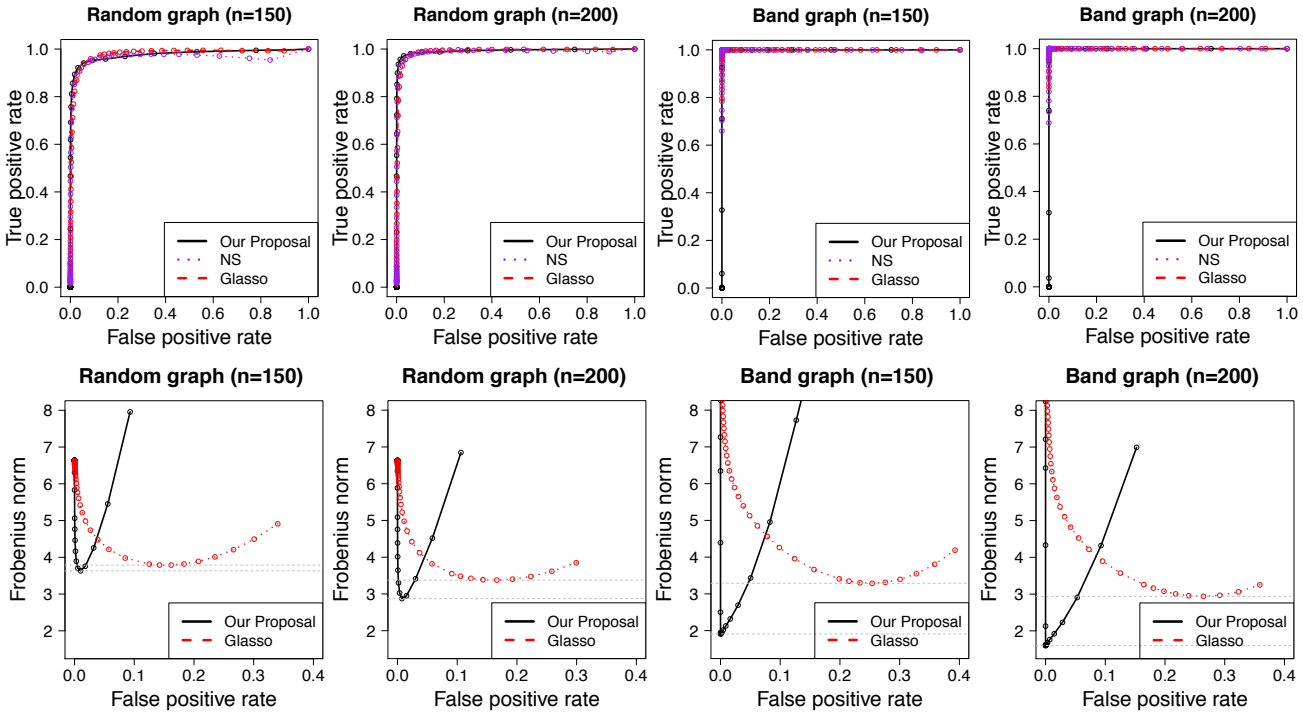

Figure 3. Row I: True and false positive rates, averaged over 100 data sets with $d = 150$, for random and band graphs, respectively. Row II: Difference between the estimated and the true inverse covariance matrices under the Frobenius norm. The different curves are obtained by varying the sparsity tuning parameter for each of the methods.

with $n = 150$, we see that the minimum error under the Frobenius norm for our proposal is smaller than that of the graphical lasso. As we increase the number of observations to $n = 200$, the difference between the minimum error for the two proposals are more apparent. More interestingly, the region for which our proposal has lower Frobenius norm than the graphical lasso is the primary region of interest. This is because an ideal estimator is one that has a low false positive rate while maintaining a high true positive rate with low error under the Frobenius norm. In contrast, the region for which the graphical lasso does better under the Frobenius norm is not the primary region of interest due to the high false positive rate. We see similar results for the band graph setting.

6. Conclusion and Discussions

We propose the graphical nonconvex optimization, which we approximate via a sequence of convex programs, for estimating the inverse correlation and concentration matrices. We prove that our proposed estimators have better statistical rates of convergence compared to existing approaches. The proposed method is sequential convex in nature and thus is computationally tractable. Yet surprisingly, it produces estimators with oracle rate of convergence as if the global optimum for the penalized nonconvex problem could be obtained. Our results stem from the contraction property we have proven, i.e., every convex problem contracts the previous estimator by a 0.5-fraction towards the optimal rate of convergence. Roughly speaking, since the first convex program achieves rate of convergence $\sqrt{s \log d/n}$ and the optimal rate is $\sqrt{s/n}$ under the Frobenius-norm, it can be shown that we need $\log[\log(d)]$ convex programs to achieve the optimal rate of $\sqrt{s/n}$ from $\sqrt{s \log d/n}$.

Our work can be applied to many different topics: low rank matrix completion problems, high-dimensional quantile regression and many others. We conjecture that in all of the aforementioned topics, a similar sequential convex approximation can be proposed and can possibly give faster rate, with controlled computing resources. It is also interesting to see how our algorithm works in large-scale distributed systems. Are there any fundamental tradeoffs between statistical efficiency, communication and algorithmic complexity? We leave these as future research topics.

Acknowledgements

We thank all three reviewers for their insightful comments. Qiang Sun is supported by Connaught New Researcher Award, NSERC Grant RGPIN-2018-06484. Tong Zhang is supported by NSF IIS1407939.

References

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

Cai, T., Liu, W., and Luo, X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Cai, T., Ren, Z., and Zhou, H. H. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016a.

Cai, T. T., Liu, W., and Zhou, H. H. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016b.

Drton, M. and Maathuis, M. H. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2016.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Fan, J., Liu, H., Sun, Q., and Zhang, T. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814–841, 2018.

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

Ge, D., Jiang, X., and Ye, Y. A note on the complexity of lp minimization. *Mathematical Programming*, 129(2): 285–299, 2011.

Lam, C. and Fan, J. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6):4254–4278, 2009.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Loh, P.-L. and Wainwright, M. J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16: 559–616, 2015.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pp. 1436–1462, 2006.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Wang, Z., Liu, H., and Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164–2201, 2014.

Yuan, M. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

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

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942, 2010a.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010b.