Electronic Journal of Statistics
Vol. 9 (2015) 1205–1229
ISSN: 1935-7524
DOI: 10.1214/15-EJS1031

Confidence intervals for high-dimensional inverse covariance estimation

Jana Janková and Sara van de Geer

Seminar for Statistics
ETH Zürich
e-mail: jankova@stat.math.ethz.ch; geer@stat.math.ethz.ch

Abstract: We propose methodology for statistical inference for low-dimensional parameters of sparse precision matrices in a high-dimensional setting. Our method leads to a non-sparse estimator of the precision matrix whose entries have a Gaussian limiting distribution. Asymptotic properties of the novel estimator are analyzed for the case of sub-Gaussian observations under a sparsity assumption on the entries of the true precision matrix and regularity conditions. Thresholding the de-sparsified estimator gives guarantees for edge selection in the associated graphical model. Performance of the proposed method is illustrated in a simulation study.

MSC 2010 subject classifications: Primary 62J07; secondary 62F12.
Keywords and phrases: Confidence intervals, graphical Lasso, high-dimensional, precision matrix, sparsity.

Received March 2014.

1. Introduction

A large number of methods has been proposed for the problem of inverse covariance estimation in high-dimensional settings, where the number of parameters may be much larger than the sample size. Common procedures in literature typically take advantage of thresholding which leads to estimators whose asymptotic distribution largely depends on the underlying unknown parameter [16] and is in general not tractable, which makes it challenging to establish any results for statistical inference. In this paper, motivated by the semi-parametric approach adopted in [35] and [39], we propose an asymptotically normal non-sparse estimator of the precision matrix which leads to confidence regions and testing for low-dimensional parameters.

The problem of estimating the inverse covariance matrix in high dimensions naturally arises in a wide variety of application domains, such as graphical modeling of brain connectivity based on FMRI brain analysis [24], gene regulatory network discovery [29], financial data processing, social network analysis and climate data analysis. The development of methodology for high-dimensional inference is of interest for instance in differential networks, which comprise two

sample comparisons of high-dimensional graphical models where the goal is to test equality of networks corresponding to two different populations. Differential networks find application e.g. in cancer studies [30].

Consider an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^p$ of size n from a zero-mean distribution with unknown covariance matrix $\Sigma^* \in \mathbb{R}^{p \times p}$. Denoting the inverse covariance matrix, often referred to as the precision or concentration matrix, as $\Theta^* = (\Sigma^*)^{-1}$, the goal is to the estimate Θ^* in a setting where $p \gg n$. The most natural candidate for an estimator of the covariance matrix is presumably the sample covariance matrix. However, when $p > n$, the sample covariance matrix is singular with probability one. Even when p/n tends to a constant, the covariance matrix exhibits poor performance [15].

Different structural assumptions have been imposed on the model to allow for consistent estimation in the regime $p \gg n$, here we consider in particular sparsity assumptions on the number of non-zero elements of the precision matrix. Let $\mathcal{V} := \{1, \ldots, p\}$, let $S \equiv S(\Theta^*) := \{(i,j) \in \mathcal{V} \times \mathcal{V} : \Theta^*_{ij} \neq 0\}$ be the set of all non-zero entries of Θ^* and denote the cardinality of S by s. Use $S^c(\Theta^*)$ for the complement of $S(\Theta^*)$ in $\mathcal{V} \times \mathcal{V}$. We shall impose a sparsity assumption on the maximum row cardinality of Θ^*; therefore define $d = d_n$ as follows

$$d := \max_{i \in \{1,\ldots,p\}} |\{j \in \mathcal{V} : \Theta^*_{ij} \neq 0\}|.$$

Estimation of precision matrices is of interest in Gaussian graphical modeling where the entries of the precision matrix represent conditional dependences between the variables [17]. Suppose that $X = (X^1, \ldots, X^p) \sim \mathcal{N}(0, \Sigma^*)$ and associate the variables $X^1, \ldots, X^p$ with the vertex set $\mathcal{V} = \{1, \ldots, p\}$ of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with an edge set $\mathcal{E}$. A pair (i,j) is included in the edge set if and only if the variables X^i and X^j are not independent given all remaining variables. Under $X \sim \mathcal{N}(0, \Sigma^*)$, a pair of variables is conditionally independent given all remaining variables if and only if the corresponding entry in the precision matrix $\Theta^* = (\Sigma^*)^{-1}$ is zero. Hence a pair of variables is contained in the edge set if and only if the corresponding entry in the inverse covariance matrix $\Theta^* = (\Sigma^*)^{-1}$ is non-zero ([17]). Elements of the precision matrix may thus be interpreted as the edge weights in the Gaussian graphical model. The parameter d corresponds to the maximum node degree in the associated Gaussian graphical model and thus sparsity assumptions on d translate to sparsity of the edges in the graphical model.

1.1. Overview of related work

Existing work on statistical inference in high dimensional settings has mostly focused on inference for parameters in linear models and generalized linear models [35, 39, 20, 13, 9, 8, 25]. In particular we mention the paper [39] where a semi-parametric projection approach was proposed for testing and construction of confidence intervals for low-dimensional parameters. The proposed method is based on the Lasso estimator for which the Karush-Kuhn-Tucker conditions are

"inverted" to obtain a de-sparsified estimator. The approach leads to asymptotically normal and efficient (in a semi-parametric sense) estimation of the regression coefficients and an extension of the method to generalized linear models is given in [35]. The key assumption which allows for asymptotically normal estimation requires sparsity of order $\sqrt{n}/\log p$ in the high-dimensional parameter vector and the method relies on ℓ_1 norm error bound of the Lasso. The paper [13] essentially follows the same approach as [39] but uses a different approach to find an approximate inverse for the sample covariance matrix.

Further methodology for inference for the regression coefficients in high-dimensional regression includes methods based on sample splitting [22, 36], bootstrapping approach [9, 8], inference after variable selection [3] and other [20, 2].

Estimation of precision matrices is a problem closely related to linear regression and in high dimensions has been extensively studied in terms of point estimation. Less work has yet been done on inference for precision matrices in this setting. We mention the work [27] which suggests a regression approach leading to an asymptotically normal estimator for elements of the precision matrix, under row sparsity of order $\sqrt{n}/\log p$, bounded spectrum of the true precision matrix and Gaussianity of the underlying distribution. The procedure regresses each pair of variables (X^i, X^j) on all the remaining variables for each $(i, j) \in \mathcal{V} \times \mathcal{V}$ to obtain an estimate of the noise level of the conditional distribution of (X^i, X^j). This requires $\mathcal{O}(p^2)$ high-dimensional regressions with the square-root Lasso ([1]).

The large amount of work that has studied methodology for point estimation of precision matrices (a selected list includes [11, 21, 37, 6, 31, 4]) typically uses regularization in terms of ℓ_1 norm or some sort of thresholding of the sample covariance matrix. Hence they do not immediately lead to results for inference, but we show they may serve as good initial estimators to construct asymptotically normal estimators.

Here we consider in particular the graphical Lasso, which minimizes the negative Gaussian log-likelihood with regularization in terms of the ℓ_1 norm of the off-diagonal entries of the precision matrix and has been studied in detail in several papers [11, 28, 26] and [38]. The optimization problem corresponding to graphical Lasso is a convex optimization problem that can be solved with coordinate descent methods [10, 11] in polynomial time.

The asymptotic behaviour of the graphical Lasso has been studied in [28] (see also [23]) which derives rates of convergence in Frobenius norm of order $O((p + d)\log p/n)$ under mild conditions on the eigenvalues of Θ^* and under sparsity $(p + d)\log p/n \to 0$. High-dimensionality here is reflected in p being allowed to grow as a function of n, however, in limit, $p/n \to 0$ is required. The high-dimensional setting $p \gg n$ is considered in [26], where convergence rates for the supremum norm of order $O(\kappa_{\Gamma^*}\sqrt{\log p/n})$ are derived under an irrepresentability condition on the true precision matrix Θ^*, sparsity $d^2 \log p/n \to 0$ and sub-Gaussian tails of the underlying distribution. The rates depend on certain quantities κ_{Γ^*} and κ_{Σ^*}, where κ_{Γ^*} is the ℓ_1 matrix norm of the inverse of a certain subset of the Hessian matrix $\Gamma^* = \Sigma^* \otimes \Sigma^*$ and κ_{Σ^*} the ℓ_1 matrix

norm of the true covariance matrix Σ^*. For reader's convenience we discuss the results in more detail in Section 1.3 and appendix A.

Further methodology on estimation of precision matrices in particular includes the regression approach [21, 37, 6] and [31] which uses a Lasso-type algorithm or Dantzig selector [7] to estimate each column or a smaller part of the precision matrix individually, thresholding of the sample covariance matrix [4] or a combination thereof.

1.2. Outline

In this paper, we propose a de-sparsified estimator based on the graphical Lasso and study its theoretical properties for low-dimensional statistical inference and edge selection in the associated graphical model. The work closely follows the approach of [35], which builds on "inverting" the necessary Karush-Kuhn-Tucker conditions for an optimization problem. The paper [35] demonstrates this method for the case of linear regression and generalized linear models, while we apply the idea to a fully nonlinear estimator. By inverting the KKT conditions, we obtain a de-sparsified graphical Lasso estimator and consequently we analyze its asymptotic properties. Asymptotic normality of the new estimator is proved for sub-Gaussian observations, under regularity conditions on the true precision matrix Θ^*. The estimator may be thresholded again to give guarantees for edge selection in the associated graphical model. The performance of the method is illustrated on both simulated and real data.

The paper is organized as follows. In Section 1.3, we briefly introduce the model. Section 2 contains the main results. Section 3 illustrates the theoretical results in a simulation study and on a real data set. Finally, Section 4 contains proofs.

Notation. For two matrices A and B, use $A \otimes B$ to denote the Kronecker product of A and B. For a vector $x \in \mathbb{R}^d$ and $p \in (0, \infty]$ we use the notation $\|x\|_p$ to denote the $p-$norm of x in the classical sense. For a matrix $A \in \mathbb{R}^{d \times d}$ we use the notations $|||A|||_\infty = \max_i \|e_i^T A\|_1$, $|||A|||_1 = |||A^T|||_\infty$ and $\|A\|_\infty = \max_{i,j} |A_{ij}|$. The symbol $\mathrm{vec}(A)$ denotes the vectorized version of a matrix A obtained by stacking rows of A on each other. By e_i we denote a p-dimensional vector of zeros with one at position i and by $e_{ij} := e_i \otimes e_j$ a p^2-dimensional vector of zeros with one at position indexed by (i, j).

For sequences f_n, g_n, we write $f_n = O(g_n)$ if $|f_n| \leq C|g_n|$ for some $C > 0$ independent of n and all $n > C$. Analogously, we write $f_n = \Omega(g_n)$ if $|f_n| \geq C|g_n|$ for some $C > 0$ independent of n and all $n > C$. We write $f_n \asymp g_n$ if both $f_n = \mathcal{O}(g_n)$ and $f_n = \Omega(g_n)$ hold. Finally, $f_n = o(g_n)$ if $\lim_{n\to\infty} f_n/g_n = 0$.

We use S_+^p to denote the cone of positive semi-definite $p \times p$ matrices, i.e. $S_+^p := \{A \in \mathbb{R}^{p\times p} | A = A^T, A \succeq 0\}$ and S_{++}^p to denote the set of positive definite $p \times p$ matrices, $S_{++}^p := \{A \in \mathbb{R}^{p\times p} | A = A^T, A \succ 0\}$.

The components of a vector $X \in \mathbb{R}^p$ will be denoted by upper indices, i.e. $X = (X^1, \ldots, X^p)$. Elements of matrices will be typically denoted by lower indices, e.g. A_{ij}. We use $\rightsquigarrow$ to denote convergence in distribution.

1.3. Model setup

Definition 1. A real zero-mean random variable X is sub-Gaussian if there exists $K > 0$ such that

$$\mathbb{E}e^{X^2/K^2} \leq 2. \tag{1}$$

Condition (C1) implies a bound on the moment generating function $\mathbb{E}^{tX} \leq e^{\frac{3}{2}K^2t^2}$ for all $t > 0$ and a tail bound $\mathbb{P}(|X| > t) \leq 2e^{-\frac{t}{6K^2}}$ for all $t > 0$, which are both equivalent characterizations of sub-Gaussianity. A prime example of a sub-Gaussian random variable is a zero-mean Gaussian random variable.

We shall consider the following sub-Gaussianity conditions for random vectors $X = (X^1, \ldots, X^p)$ with zero mean and covariance matrix Σ^*.

Condition (C1) (Sub-Gaussianity condition)**.** All normalized components $X^i/\sqrt{\Sigma^*_{ii}}, i = 1, \ldots, p$ of the zero-mean random vector $X = (X^1, \ldots, X^p)$ with covariance matrix Σ^* are sub-Gaussian random variables with a common parameter $K > 0$.

The condition (C1) is weaker than requiring the sub-Gaussianity of the whole vector $X = (X^1, \ldots, X^p)$ in the following sense.

Condition (C2) (Sub-Gaussianity vector condition)**.** A zero-mean random vector $X \in \mathbb{R}^p$ satisfies the sub-Gaussianity vector condition if there exists a constant $K > 0$ such that

$$\sup_{\alpha \in \mathbb{R}^p : \|\alpha\|_2 \leq 1} \mathbb{E}e^{|\alpha^T X|^2/K^2} \leq 2. \tag{2}$$

If a random vector $X = (X^1, \ldots, X^p)$ satisfies (C2) with a constant K, then each component X^i satisfies (1) with K.

We now review some notation and results related to the graphical Lasso estimator [11] on which our further analysis is based. Consider an i.i.d. sample $X_1, \ldots, X_n$ distributed as X with $\mathbb{E}X = 0, \operatorname{cov}(X) = \Sigma^*$. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ be the sample covariance matrix. We further write $\hat{\Sigma}_{ij} := (\hat{\Sigma})_{ij}$ for the (i,j)-th element of $\hat{\Sigma}$, $(i,j) \in \mathcal{V} \times \mathcal{V}$. The graphical Lasso estimator $\hat{\Theta}$ [11] is defined as the solution to the optimization problem

$$\hat{\Theta} := \arg\min_{\Theta \in S^p_{++}} \left\{ \operatorname{trace}(\Theta^T \hat{\Sigma}) - \log\det(\Theta) + \lambda \|\Theta\|_{1,\text{off}} \right\}, \tag{P1}$$

where $\hat{\Sigma}$ is the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ and $\|\cdot\|_{1,\text{off}}$ is the ℓ_1 off-diagonal penalty, $\|\Theta\|_{1,\text{off}} = \sum_{i \neq j} |\Theta_{ij}|$. When the data is normally distributed, (P1) is equivalent to ℓ_1-penalized maximum likelihood for the precision matrix.

In our analysis, we rely on the results on rates of convergence derived in [26] for the graphical Lasso in supremum norm. The work [26] assumes an irrepresentability condition which is a rather restrictive condition in the linear regression setting [34]. However, other literature on the graphical Lasso [38]

likewise assumes irrepresentability condition or otherwise assumes $p/n \to 0$ [28]. In the linear regression setting, irrepresentable conditions are sufficient for variable selection [34, 21, 40].

The analysis of rates of convergence of the graphical Lasso in [26] in addition considers certain functions of the true precision matrix Θ^*, which we now define.

Let κ_{Σ^*} be the ℓ_∞ operator norm of the true covariance matrix Σ^*, i.e.

$$\kappa_{\Sigma^*} = |||\Sigma^*|||_\infty = \max_i \sum_{j=1}^p |\Sigma^*_{ij}|.$$

The parameter κ_{Σ^*} then measures the size of entries in Σ^*.

Example 1. Consider the Töplitz matrix $\Sigma^*_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, p$, where $|\rho| < 1$. Then $\kappa_{\Sigma^*} = (1-\rho^p)/(1-\rho) = \mathcal{O}(1)$ if ρ is bounded away from 1.

We next consider the Hessian $\Gamma(\Theta)$ of the negative log-likelihood function $\ell(\Theta) = \operatorname{tr}(\Theta^T\hat{\Sigma}) - \log\det(\Theta)$. The entries of the gradient of ℓ are given by ([12])

$$\frac{\partial \ell(\Theta)}{\partial \Theta_{ij}} = \hat{\Sigma}_{ij} - (\Theta^{-1})_{ij}.$$

The Hessian matrix is then indexed by pairs of edges $((i,j),(k,l))$ and the $((i,j),(k,l))$-th entry takes the form

$$\frac{\partial^2 \ell(\Theta)}{\partial \Theta_{kl}\partial \Theta_{ij}} = \frac{\partial(\hat{\Sigma}_{ij} - (\Theta^{-1})_{ij})}{\partial \Theta_{kl}\partial \Theta_{ij}} = e_i^T\Theta^{-1}e_k e_l^T\Theta^{-1}e_j = \Sigma_{ik}\Sigma_{lj},$$

where $\Sigma = \Theta^{-1}$. In matrix form, we obtain

$$\Gamma(\Theta) = \Sigma \otimes \Sigma.$$

By (i,j)-th column of $\Sigma \otimes \Sigma$ we refer to the $p^2 \times 1$ vector $\Sigma \otimes \Sigma \operatorname{vec}(e_i e_j^T)$ and (i,j)-th row of $\Sigma \otimes \Sigma$ is its transpose. The (i,j)-th row of $\Sigma \otimes \Sigma$ contains all mixed partial derivatives of ℓ with respect to Θ_{ij} and Θ_{kl} where $k, l = 1, \ldots, p$. Note that $\Theta \otimes \Theta$ may be viewed as a four-dimensional tensor.

We impose some restrictions on the Hessian Γ evaluated at the true Θ^*, $\Gamma^* := \Gamma(\Theta^*)$. To this end, let us fix the following notation. For any two subsets T and T' of $\mathcal{V} \times \mathcal{V}$, we use $\Gamma^*_{TT'}$ to denote the $|T| \times |T'|$ matrix with rows and columns of Γ^* indexed by T and T' respectively.

Consequently, define κ_{Γ^*} to be the ℓ_∞ operator norm of the inverse of the matrix

$$\Gamma^*_{SS} = [\Sigma^* \otimes \Sigma^*]_{SS} \in \mathbb{R}^{s \times s}.$$

i.e., $\kappa_{\Gamma^*} = \left|\left|\left|(\Gamma^*_{SS})^{-1}\right|\right|\right|_\infty$.

The parameter κ_{Γ^*} then measures the size of entries in Θ^* and assumptions on its growth are similar to sparsity assumptions on Θ^*.

Example 2. The parameter κ_{Γ^*} is difficult to track in general as it involves inversion of a certain sub-matrix of the Hessian. A tractable example is the situation when Θ^* is a block diagonal matrix with blocks $B_1, \ldots, B_k$ for some $1 \leq k \leq p$ which only contain non-zero (although possibly arbitrarily small) entries and the remaining off-diagonal entries of Θ^* are zero. Suppose that the sizes of the blocks are $b_1, \ldots, b_k$ and denote $d := \max_{i=1,\ldots,k} b_i$. This corresponds to a graph with k completely connected but mutually isolated subgraphs with maximum vertex degree d. Using that block matrices can be easily inverted by inverting each block separately (and using that $(A \otimes A)^{-1} = A^{-1} \otimes A^{-1}$), some calculations give

$$\kappa_{\Gamma^*} = \max_{i=1,\ldots,k} |||B_i \otimes B_i|||_\infty = \max_{i=1,\ldots,k} |||B_i|||_\infty^2.$$

The size of κ_{Γ^*} thus depends on the size of entries in $B_i's$. We clearly have the upper bound

$$\kappa_{\Gamma^*} \leq d \max_{i=1,\ldots,k} \max_{j=1,\ldots,b_i} (B_i^{jj})^2 \leq d\Lambda_{\max}^2(\Theta^*).$$

Hence if the maximum eigenvalue of Θ^* is bounded, then $\kappa_{\Gamma^*} = \mathcal{O}(d)$. This bound is attained for instance when all entries in some row of the block of size $d \times d$ were bounded away from zero uniformly in n and $\Lambda_{\max}(\Theta^*) = \mathcal{O}(1)$.

In the trivial case when Σ^* and Θ^* are diagonal matrices, we have $S = \{(i,i) : i = 1, \ldots, p\}$ and and $\kappa_{\Gamma^*} = \max_i (\Theta_{ii}^*)^2$. Then clearly $\kappa_{\Gamma^*} \leq \Lambda_{\max}(\Theta^*)$.

The most difficult situation is when Θ^* a single block, so $S = \{(i,j) : i,j = 1, \ldots, p\}$ and $\kappa_{\Gamma^*} = |||(\Sigma^* \otimes \Sigma^*)^{-1}|||_\infty = |||\Theta^* \otimes \Theta^*|||_\infty = |||\Theta^*|||_\infty^2$. If the entries in Θ^* are fast decaying, for instance $\Theta_{ij}^* = \tau^{|i-j|}$, $|\tau| < 1$ then

$$\kappa_{\Gamma^*} = (1-\tau^p)^2/(1-\tau)^2 = \mathcal{O}(1). \qquad \square$$

Assumption (A1) (Irrepresentability condition)**.** *There exists* $\alpha \in (0,1]$ *such that*

$$\max_{e \in S^c} \|\Gamma_{eS}^*(\Gamma_{SS}^*)^{-1}\|_1 \leq 1 - \alpha. \tag{3}$$

Condition (A1) is an analogy of the irrepresentable condition for variable selection in linear regression [34]. If we define the zero-mean edge random variables ([26]) as

$$Y_{(i,j)} := X_i X_j - \mathbb{E}(X_i X_j),$$

then the matrix Γ^* corresponds to covariances of the edge variables, in particular $\Gamma^*_{(i,j),(k,l)} + \Gamma^*_{(j,i),(k,l)} = \mathrm{cov}(Y_{(i,j)}, Y_{(k,l)})$. The interpretation of (A1) is that we require that no edge variable $Y_{(j,k)}$ which is not included in the edge set S is highly correlated with variables in the edge set [26]. The parameter α then is a measure of this correlation with the correlation growing when $\alpha \to 0$.

Note that one may view Γ^* as a four-dimensional tensor, and the irrepresentability condition is then imposed on the sub-blocks of this tensor.

Remark 1. The results obtained in [26] (see also Lemma 9) imply that under the irrepresentability condition (A1), the model $\hat{S}$ selected by the graphical

Lasso satisfies $\hat{S} \subseteq S$ with high probability. Moreover, under a beta-min condition on the entries of the true precision matrix, Lemma 9 part (b) implies exact variable selection, i.e. $\hat{S} = S$ with high probability. Several works then suggest to use post-model selection methods, by which we refer to the two-step procedure resulting from first selecting a model and then estimating the parameters in the selected model (e.g. by maximum likelihood). For estimation of regression coefficients in the linear model, simple post-model selection methods have been proposed e.g. in [14, 7]. Other approaches using post-model selection in a more involved way include e.g. [2, 3]. We mention that several concerns have been raised considering simple post-model selection methods, which are elaborated on in the papers [18, 19] or [3]. The procedure we suggest in the present paper in principle does not rely on model selection (see also Remark 4). The advantage of our procedure over post-model selection methods is likely to arise in situations when there are small but non-zero parameters and thus the beta-min type condition which guarantees exact variable selection is violated.

Assumption (A2) (Bounded eigenvalues)**.** *There exists* $L \asymp 1$ *such that*

$$1/L \leq \Lambda_{\min}(\Theta^*) \leq \Lambda_{\max}(\Theta^*) \leq L.$$

Remark 2. In our analysis to follow in Section 2, we keep track of the quantities κ_{Σ^*} and κ_{Γ^*} defined above and they appear in the main result (Theorem 1). Some examples where the behaviour of κ_{Σ^*} and κ_{Γ^*} is tractable were discussed in Examples 1 and 2. An example of a situation when κ_{Σ^*} is bounded is for instance the Töplitz covariance structure. It is not easy to see when κ_{Γ^*} is bounded, as it involves inversion of a certain sub-matrix of the Hessian. This is of similar difficulty as verification of the irrepresentability condition (see Assumption (A1)), which is typically considered only on small examples [26] ($p = 4$), [20] ($p = 4$).

2. Main results

In this Section we present the main results which imply inference for individual parameters of the precision matrix. We suggest a way to modify the graphical Lasso estimator by removing the bias term associated with the penalty. To this end we consider the Karush-Kuhn-Tucker (KKT) conditions for the graphical Lasso. For any $\lambda_n > 0$ and $\hat{\Sigma}$ with strictly positive diagonal elements, the optimization problem (P1) has a unique solution $\hat{\Theta}_n \in S^p_{++}$ which is characterized by the KKT conditions

$$\hat{\Sigma} - \hat{\Theta}^{-1} + \lambda \hat{Z} = 0, \tag{4}$$

where the matrix $\hat{Z}$ belongs to the sub-differential of the off-diagonal norm $\|\cdot\|_{1,\text{off}}$ evaluated at $\hat{\Theta}$ (Lemma 3 in [26]).

First we "invert" the KKT conditions (4) by multiplying them by the inverse of the Hessian of the negative log-likelihood, i.e. $(\Gamma^*)^{-1} = (\Sigma^* \otimes \Sigma^*)^{-1} = (\Sigma^*)^{-1} \otimes (\Sigma^*)^{-1} = \Theta^* \otimes \Theta^*$ which may be approximated by plugging in the graphical Lasso estimator to obtain $\hat{\Theta} \otimes \hat{\Theta}$. As noted in Section 1.3, when

$X_1, \ldots, X_n$ are Gaussian, there is a correspondence between $\Sigma^* \otimes \Sigma^*$ and the Fisher information matrix for Θ^*.

By the properties of the Kronecker product [12], this is equivalent to multiplication of (4) by $\hat{\Theta}$ from left and right.

$$\hat{\Theta}\hat{\Sigma}\hat{\Theta} - \hat{\Theta} + \hat{\Theta}\lambda\hat{Z}\hat{\Theta} = 0.$$

Denoting $W := \hat{\Sigma} - \Sigma^*$ and rearranging yields

$$\hat{\Theta} + \hat{\Theta}\lambda\hat{Z}\hat{\Theta} - \Theta^* = -\Theta^* W \Theta^* + \text{rem}, \tag{5}$$

where

$$\text{rem} := -(\hat{\Theta} - \Theta^*)W\Theta^* - (\hat{\Theta}\hat{\Sigma} - I)(\hat{\Theta} - \Theta^*). \tag{6}$$

The term rem is shown to be small under sufficient sparsity (Lemma 1) and the leading term $\Theta^* W \Theta^*$ is (elementwise) asymptotically normal. This suggests to take the modified non-sparse estimator $\hat{\Theta} + \hat{\Theta}\lambda\hat{Z}\hat{\Theta}$ as an estimator for Θ^*. Recall that by the KKT conditions (4), $\lambda\hat{Z}$ may be expressed as $\hat{\Theta}^{-1} - \hat{\Sigma}$. Hence define the de-sparsified graphical Lasso estimator as follows

$$\hat{T} := \hat{\Theta} + \hat{\Theta}\lambda\hat{Z}\hat{\Theta} = 2\hat{\Theta} - \hat{\Theta}\hat{\Sigma}\hat{\Theta}. \tag{7}$$

The following auxiliary Lemma gives a bound for the remainder (6) under sub-Gaussian tail assumptions (C1) and (C2).

Lemma 1. *Suppose that $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent and distributed as $X = (X^1, \ldots, X^p)$ with $\mathbb{E}X = 0$, $cov(X) = \Sigma^*$. Let $\Theta^* = (\Sigma^*)^{-1}$ exist and satisfy the irrepresentability condition (A1) with a constant $\alpha \in (0, 1]$. Let $\hat{\Theta}$ be the solution to the optimization problem (P1) with tuning parameter $\lambda_n = \frac{8}{\alpha}\delta_n$ where for some $\gamma > 2$,*

$$\delta_n := 8(1 + 12K^2) \max_i \Sigma^*_{ii} \sqrt{2\frac{\log(4p^\gamma)}{n}},$$

where K is specified below. Suppose that the sparsity assumption

$$d \leq \frac{1}{6(1 + 8/\alpha)\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}\delta_n}$$

and assumption (A2) are satisfied.

(i) Suppose that $X_1, \ldots, X_n \in \mathbb{R}^p$ satisfy (C1) with $K = \mathcal{O}(1)$. Then it follows that

$$\|\text{rem}\|_\infty = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\alpha^2}\kappa_{\Gamma^*}\max\{d^{3/2}\log p/n, \frac{1}{\alpha}\kappa_{\Gamma^*}d^2(\log p/n)^{3/2}\}\right).$$

(ii) Suppose that $X_1, \ldots, X_n \in \mathbb{R}^p$ satisfy (C2) with $K = \mathcal{O}(1)$. Then

$$\|\text{rem}\|_\infty = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\alpha^2}\kappa^2_{\Gamma^*}\kappa_{\Sigma^*}d\log p/n\right).$$

The quantities $\kappa_{\Sigma^*}, \kappa_{\Gamma^*}$ involved in Lemma 1 measure the size of entries in Σ^* and $(\Gamma^*_{SS})^{-1}$ as discussed in Section 1.3. The parameter α corresponds to the irrepresentability condition (A1) and affects the rates when it approaches zero.

If we assume the quantities in Lemma 1 are bounded, i.e. $1/\alpha = \mathcal{O}(1)$, $\kappa_{\Sigma^*} = \mathcal{O}(1)$ and $\kappa_{\Gamma^*} = \mathcal{O}(1)$ then the sparsity assumption reduces to

$$d \leq \sqrt{n/\log p}$$

and under (C1) we have

$$\|\text{rem}\|_\infty = \mathcal{O}_{\mathbb{P}}\left(d^{\frac{3}{2}}\frac{\log p}{\sqrt{n}}\right),$$

under (C2) we have

$$\|\text{rem}\|_\infty = \mathcal{O}_{\mathbb{P}}\left(d\frac{\log p}{\sqrt{n}}\right).$$

Consequently, we establish asymptotic normality of each element $\hat{T}_{ij}$ of the de-sparsified estimator in Theorem 1 below. Since our aim is inference about individual elements of Θ^*, fix $(i,j) \in \mathcal{V} \times \mathcal{V}$. The k-th column of Θ^* will be denoted by $\Theta^*_k \in \mathbb{R}^p$, $k = 1, \ldots, p$.

Theorem 1. *Suppose that $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent and distributed as $X = (X^1, \ldots, X^p)$ with $\mathbb{E}X = 0$, $cov(X) = \Sigma^*$. Let $\Theta^* = (\Sigma^*)^{-1}$ exist, satisfy the irrepresentability condition (A1) with a constant $\alpha \in (0,1]$ and assumption (A2). Let*

$$\sigma^2_{ij} := \text{Var}({\Theta^*_i}^T X_1 X_1^T \Theta^*_j)$$

and suppose that $1/\sigma_{ij} = \mathcal{O}(1)$. Suppose that $\hat{\Theta}$ is the solution to the optimization problem (P1) with tuning parameter $\lambda_n \asymp \sqrt{\log p/n}$. Suppose the sparsity assumption under (C1)

$$d^{3/2} = o\left(\frac{\sqrt{n}}{C_1 \log p}\right), \tag{8}$$

where

$$C_1 := \max\left\{\frac{\kappa_{\Gamma^*}}{\alpha^2}, \frac{\kappa^2_{\Gamma^*}}{\alpha^{9/8}} n^{-1/4}(\log p)^{1/8}, \frac{\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}^{3/2}}{\alpha^{3/2}}(n\log p)^{-1/4}\right\}.$$

and under (C2)

$$d = o\left(\frac{\sqrt{n}}{C_2 \log p}\right), \tag{9}$$

where

$$C_2 := \frac{1}{\alpha}\kappa_{\Gamma^*} \max\left\{\frac{1}{\alpha}\kappa_{\Gamma^*}\kappa_{\Sigma^*}, \kappa_{\Sigma^*}(\log p)^{-1/2}, \kappa^3_{\Sigma^*}\kappa_{\Gamma^*}(\log p)^{-1/2}\right\},$$

is satisfied. Let $\hat{T}$ be the de-sparsified graphical Lasso estimator defined in (7). Then under (C1) with sparsity (8) or under (C2) with sparsity (9) for all $(i,j) \in \mathcal{V} \times \mathcal{V}$, it holds that

$$\sqrt{n}(\hat{T}_{ij} - \Theta^*_{ij})/\sigma_{ij} = Z^n_{ij} + o_{\mathbb{P}}(1), \tag{10}$$

where Z^n_{ij} converges weakly to $\mathcal{N}(0,1)$.

When the quantities κ_{Γ^*}, κ_{Σ^*} and $1/\alpha$ are assumed to be bounded, then the sparsity assumptions of Theorem 1 reduce to $d^{3/2} = o(\sqrt{n}/\log p)$ under (C1) and $d = o(\sqrt{n}/\log p)$ under (C2). The latter condition $d = o(\sqrt{n}/\log p)$ is the same sparsity assumption as required for construction of confidence intervals for regression coefficients using the de-sparsified Lasso [35].

The asymptotic variance σ_{ij} in Theorem 1 is typically unknown, so to construct confidence intervals one needs to use a consistent estimator $\hat{\sigma}_{ij} > 0$ for σ_{ij}. For the case of Gaussian observations, we may easily calculate the theoretical variance and plug in the estimate $\hat{\Theta}$ in place of the unknown Θ^* as is displayed in Lemma 2 below.

Lemma 2. *Suppose that assumption (A2) is satisfied and assume that $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent $\mathcal{N}(0, \Sigma^*)$. Let $\hat{\Theta}$ be the graphical Lasso estimator, let $\lambda \asymp \sqrt{\log p/n}$ and suppose the sparsity assumption*

$$d \leq \frac{\sqrt{n}}{\log p(1+8/\alpha)\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}}.$$

Then $\sigma^2_{ij} = \Theta^*_{ii}\Theta^*_{jj} + {\Theta^*_{ij}}^2$, $1/\sigma_{ij} = \mathcal{O}(1)$ and for $\hat{\sigma}^2_{ij} := \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}^2_{ij}$ we have

$$|\hat{\sigma}^2_{ij} - \sigma^2_{ij}| = \mathcal{O}_{\mathbb{P}}\left(1/\alpha\kappa_{\Gamma^*}\sqrt{\log p/n}\right).$$

Hence by Lemma 2 under assumptions of Theorem 1 we have $\sigma_{ij}/\hat{\sigma}_{ij} = o_{\mathbb{P}}(1)$ and we may replace σ_{ij} by $\hat{\sigma}_{ij}$.

Theorem 1 also implies convergence rates for the de-sparsified graphical Lasso estimator $\hat{T}$ in supremum norm. Under (C2) and assumptions of Theorem 1 we have the upper bound

$$\begin{aligned} \|\hat{T} - \Theta^*\|_\infty &\leq \|\Theta^* W \Theta^*\|_\infty + \|\mathrm{rem}\|_\infty \\ &= \mathcal{O}_{\mathbb{P}}\left(\max\left\{\sqrt{\frac{\log p}{n}}, \frac{1}{\alpha^2}\kappa^2_{\Gamma^*}\kappa_{\Sigma^*}\frac{\log p}{n}\right\}\right). \end{aligned} \tag{11}$$

Assuming κ_{Γ^*}, κ_{Σ^*} and $1/\alpha$ bounded implies

$$\|\hat{T} - \Theta^*\|_\infty = \mathcal{O}_{\mathbb{P}}(\max\{\sqrt{\log p/n}, d\log p/n\}). \tag{12}$$

Under sparsity $d = o(\sqrt{n}/\log p)$ we have $\|\hat{T} - \Theta^*\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n})$.

Consequently, under the conditions of Lemma 2 thresholding $\hat{T}_{ij}$ at level $\Phi^{-1}(1 - \frac{\alpha}{p(p-1)})\frac{\hat{\sigma}_{ij}}{\sqrt{n}}$ for all i, j will remove all zero entries with probability $1-\alpha$ asymptotically.

Remark 3. The quantities $\kappa_{\Sigma^*}, \kappa_{\Gamma^*}$ and α involved in our analysis arise from the deterministic analysis of the graphical Lasso as carried out in [26]. Provided that the quantities $\kappa_{\Sigma^*}, \kappa_{\Gamma^*}$ and $1/\alpha$ remain bounded and assuming sub-Gaussianity (C2), the only additional restriction that arises from our analysis is the sparsity restriction by a factor $\sqrt{n}$ which is needed to ensure that the remainder term in Lemma 1 vanishes asymptotically. As mentioned above, this is the same assumption which is needed to establish asymptotic normality of the de-sparsified Lasso in linear regression [35]. Under these assumptions, the estimator $\hat{T}$ achieves optimal rate of convergence under the assumed model which follows from (12) and the work [27].

Remark 4. One could consider the approach presented above for other initial estimators of the precision matrix than the graphical Lasso. An estimator of the precision matrix based on the nodewise regression approach ([20]) is outlined in [32]. Asymptotic normality of the estimator in [32] may then be obtained under bounded eigenvalues of the true precision matrix, row sparsity of Θ^* of small order $\sqrt{n}/\log p$ and assuming fourth-order moment conditions on the X_i's (see [32]). To avoid digressions we do not elaborate on this alternative approach in the present paper. We note that the present analysis using the graphical Lasso requires in addition to the conditions mentioned above the irrepresentability condition. However, inspection of the proof of Theorem 1 reveals that it is rather the ℓ_1-norm oracle rates that are needed, but only results assuming the irrepresentability condition are available in the literature on the graphical Lasso at the moment. It is as of yet not clear whether the irrepresentability condition is also necessary for obtaining oracle rates for the graphical Lasso. We also refer here to [33] where the results of [26] are extended using an irrepresentable condition on the small (not necessarily zero) entries of Θ^*.

3. Empirical Results

3.1. Simulation Study

In this part we illustrate the theoretical results on simulated data and demonstrate the performance of the proposed estimator on inference, giving a comparison to some alternative methodologies. To this end, we consider sparse Gaussian graphical models which may be fully specified by a precision matrix Θ^*. Thus the random sample is distributed as $X \sim \mathcal{N}(0, \Sigma^*)$, where $\Theta^* = (\Sigma^*)^{-1}$.

We consider the chain graph on p vertices where the maximum vertex degree is by definition restricted to $d = 2$. The corresponding precision matrix Θ^* is a tridiagonal matrix, $\Theta^* = \text{tridiag}(\rho, 1, \rho)$ for a given $\rho > 0$. The cardinality of the active set is then $3p-2$. To solve the graphical Lasso program (P1), we have used the implemented procedure `glasso` of Friedman et al. [11].

Denote by $\hat{\Theta}$ the graphical Lasso estimator and by $\hat{T}$ the de-sparsified graphical Lasso estimator. In figure 1 we report histograms of $\sqrt{n}(\hat{T} - \Theta^*_{ij})/\hat{\sigma}_{ij}$ for $(i, j) \in \{(1, 1), (1, 2), (1, 3), (1, 4)\}$ with the density of $\mathcal{N}(0, 1)$ superimposed. By Theorem 1, it follows that the $(1 - \alpha)100\%$ asymptotic confidence interval for

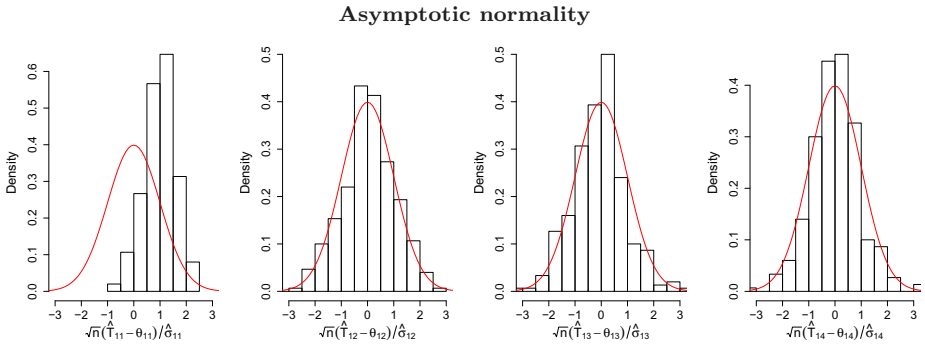

FIG 1. *Histograms for* $\sqrt{n}(\hat{T}_{ij} - \Theta^*_{ij})/\hat{\sigma}_{ij}$, $(i,j) \in \{(1,1),(1,2),(1,3),(1,4)\}$. *The sample size was* $n = 500$ *and the number of parameters* $p = 100$. *The de-sparsified graphical Lasso estimator was calculated* 300 *times. The model was the chain graph with* $\rho = 0.3$.

Θ^*_{ij} is given by

$$I_{ij} \equiv I_{ij}(\hat{\Theta}_{ij}, \alpha, n) := [\hat{T}_{ij} - \Phi^{-1}(1-\alpha/2)\frac{\sigma_{ij}}{\sqrt{n}}, \hat{T}_{ij} + \Phi^{-1}(1-\alpha/2)\frac{\sigma_{ij}}{\sqrt{n}}],$$

where we replace the unknown variance σ^2_{ij} by the plug-in estimate $\hat{\sigma}^2_{ij} = \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}^2_{ij}$ (Lemma 2). For each parameter Θ^*_{ij}, the probability that the true value Θ^*_{ij} is covered by the confidence interval was estimated by its empirical version, $\hat{\alpha}_{ij} := \mathbb{P}_N \mathbf{1}_{\{\Theta^*_{ij} \in I_{ij,\alpha}\}}$. The number of iterations used to calculate the estimates $\hat{\alpha}_{ij}$ was set to $N = 50$. Next for a set $A \subset \mathcal{V} \times \mathcal{V}$ define the average coverage over the set A as

$$\text{Avgcov}_A := \frac{1}{|A|} \sum_{(i,j)\in A} \hat{\alpha}_{ij}.$$

After obtaining the estimates $\hat{\alpha}_{ij}$, we have averaged them over the sets S and S^c to obtain $Avgcov_S$ and $Avgcov_{S^c}$, respectively. Similarly, we have calculated the average length of the confidence interval for each parameter Θ^*_{ij} from $N = 50$ iterations and again averaged these over the sets S and S^c to obtain $Avglength_S$ and $Avglength_{S^c}$.

We compare the performance of the graphical Lasso-based confidence intervals with three other methods. The first method is based on the oracle maximum likelihood estimator with the non-zero set S pre-specified. This method only serves as a theoretical benchmark as it is asymptotically efficient if the true non-zero set S is known. The second method is a post-model selection method: the maximum likelihood estimator is applied to the model selected by the graphical Lasso (see Lemma 9: under the irrepresentability condition, the graphical Lasso selects a model $\hat{S} \subseteq S$). Then the confidence intervals are constructed using asymptotic normality of the maximum likelihood estimator. The third method is based on the sample covariance matrix $\hat{\Sigma}$ which is the MLE estimator for Σ^*. Its inverse is then the maximum likelihood estimator for the

TABLE 1

The tables above show a comparison of four methods for construction of confidence intervals: the de-sparsified graphical Lasso ("De-sp. graphical Lasso"), the maximum likelihood estimator with specified set S ("MLE with specified S"), the maximum likelihood estimator based on the non-zero set $\hat{S}$ selected by the graphical Lasso ("MLE based on $\hat{S}$") and an estimator based on the sample covariance matrix $\hat{\Sigma}$ ("Sample covariance"). The table shows average coverages and average lengths of the constructed confidence intervals over the sets S and S^c (where applicable). For the method "MLE based on $\hat{S}$", the reported averages are over $\hat{S} \cap S$. The regularization parameter for the graphical Lasso was chosen $\lambda = \sqrt{\frac{\log p}{n}}$ in all simulations. The true precision matrix corresponds to a chain graph with p vertices and ρ equal to $0.3, 0.4, 0.2$ for settings S1,S2,S3, respectively

Estimated coverage probabilities and lengths

S1. $p = 80, n = 250, \rho = 0.3$	S Avgcov	S Avglength	S^c Avgcov	S^c Avglength
De-sp. graphical Lasso	0.934	0.247	0.972	0.215
MLE with specified S	0.940	0.308	–	–
MLE based on $\hat{S}$	0.887	0.325	–	–
Sample covariance	0.459	0.428	0.897	0.367

S2. $p = 100, n = 200, \rho = 0.4$	S Avgcov	S Avglength	S^c Avgcov	S^c Avglength
De-sp. graphical Lasso	0.925	0.288	0.974	0.250
MLE with specified S	0.945	0.349	–	–
MLE based on $\hat{S}$	0.856	0.374	–	–
Sample covariance	–	–	–	–

S3. $p = 100, n = 200, \rho = 0.2$	S Avgcov	S Avglength	S^c Avgcov	S^c Avglength
De-sp. graphical Lasso	0.951	0.301	0.964	0.263
MLE with specified S	0.943	0.357	–	–
MLE based on $\hat{S}$	0.747	0.328	–	–
Sample covariance	–	–	–	–

precision matrix Θ^*. In the fixed p setting, the inverse sample covariance matrix $\hat{\Sigma}^{-1}$ is thus an asymptotically normal and efficient estimator of the precision matrix when the observations are Gaussian. This allows for construction of confidence intervals in the classical way, using asymptotic normality of $\hat{\Theta} = \hat{\Sigma}^{-1}$, hence the confidence interval for Θ^*_{ij} is given by $\hat{\Theta}_{ij} \pm \Phi^{-1}(1 - \alpha/2)\frac{\hat{\sigma}_{ij}}{\sqrt{n}}$.

The results are reported in Tables 1 and 2. The de-sparsified graphical Lasso performs well also when compared to the oracle ("MLE with specified S"). In Table 1, settings S2 and S3 differ only in the value of ρ. The de-sparsified graphical Lasso performs well for both of these settings, while the the post-model selection method ("MLE based on $\hat{S}$") shows lower coverage for setting S3, where ρ is comparable in magnitude to the noise level.

In table 2, the sample size is kept fixed at $n = 100$ while the dimension of the parameter is increased, hence we observe lower coverage on S for large values of p, as expected.

TABLE 2

A table showing the performance of the de-sparsified graphical Lasso for number of parameters p taking values $100, 200, 300, 500$ *and* $n = 100$*. The regularization parameter was chosen* $\lambda = \sqrt{\frac{\log p}{n}}$ *in all simulations. The constant* ρ *is* 0.3 *in the definition of* Θ^*

Estimated coverage probabilities and lengths

Chain graph	S Avgcov	S Avglength	S^c Avgcov	S^c Avglength
p = 100	0.931	0.401	0.978	0.348
p = 200	0.917	0.400	0.984	0.349
p = 300	0.893	0.401	0.988	0.349
p = 500	0.832	0.401	0.988	0.350

The choice of the regularization parameter Theory implies that the correct choice of the regularization parameter satisfies $\lambda_n \asymp \sqrt{\frac{\log p}{n}}$. Lemma 1 gives an explicit prescription for λ, however, this theoretically obtained value of λ is very large. For all the numerical experiments, we have chosen $\lambda_n = \sqrt{\frac{\log p}{n}}$.

3.2. Real data experiment

We consider a dataset about riboflavin (vitamin B_2) production by bacillus subtilis without the response variable. The dataset is available from the `R` package `hdi`.

The dataset contains observations of $p = 4088$ logarithms of gene expression levels from $n = 71$ genetically engineered mutants of bacillus subtilis. We are interested in modeling the conditional independence structure of the covariates (logarithms of gene expression levels) and we try to estimate the associated graphical model using the de-sparsified graphical Lasso. We only consider the first 500 covariates which have the highest variances.

In the first step, we split the sample and use 10 randomly chosen observations to estimate the variances of the 500 variables. With the estimated variances, we scale the design matrix containing the remaining 61 observations.

We calculate the graphical Lasso using the tuning parameter as in the simulations, $\lambda = \sqrt{\log p/n}$, and hence calculate the de-sparsified graphical Lasso.

We threshold the de-sparsified graphical Lasso at level $\Phi^{-1}(1-\frac{\alpha}{p(p-1)})\hat{\sigma}_{ij}/\sqrt{n}$, where $\alpha = 0.05$ and $\hat{\sigma}_{ij}^2 = \hat{\Theta}_{ii}\hat{\Theta}_{ii} + \hat{\Theta}_{ij}^2$ is an estimate of the asymptotic variance calculated under the assumption of normality and using the graphical Lasso estimator $\hat{\Theta}$. We identify 5 edges as significant.

For independently permuted variables (for each variable, a different permutation is used), the conditional dependencies are broken (the truth is the empty graph), and the de-sparsified graphical Lasso correctly detects zero edges.

4. Proofs

Lemma 3. *Suppose that $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent and distributed as $X = (X^1, \ldots, X^p)$ with $\mathbb{E}X = 0$, $cov(X) = \Sigma^*$. Let $\Theta^* = (\Sigma^*)^{-1}$ exist and satisfy the irrepresentability condition (A1) with a constant $\alpha \in (0, 1]$. Let $\hat{\Theta}_n$ be the solution to the optimization problem (P1) with tuning parameter $\lambda_n = \frac{8}{\alpha}\delta_n$ for some $\delta_n > 0$ and suppose that the sparsity assumption*

$$d \leq \frac{1}{6(1+8/\alpha)\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}\delta_n} \tag{13}$$

is satisfied. Then on the set $\mathcal{T}_n = \{\|\hat{\Sigma} - \Sigma^*\|_\infty < \delta_n\}$, we have

Bound I

$$\|\text{rem}\|_\infty = \mathcal{O}\left(\frac{1}{\alpha}\kappa_{\Gamma^*}\max\{d\delta_n\|\Theta^* W\|_\infty, \frac{1}{\alpha}\kappa_{\Gamma^*}d^2\delta_n^3, \frac{1}{\alpha}\kappa_{\Gamma^*}\kappa_{\Sigma^*}d\delta_n^2\}\right) \tag{14}$$

Bound II

$$\|\text{rem}\|_\infty = \mathcal{O}\left(\frac{1}{\alpha}\kappa_{\Gamma^*}\max\{d\delta_n\|\Theta^* W\|_\infty, \frac{1}{\alpha^2}\kappa_{\Gamma^*}d^2\delta_n^3, \frac{1}{\alpha}\Lambda_{\max}(\Theta^*)d^{3/2}\delta_n^2\}\right). \tag{15}$$

Proof of Lemma 3. The KKT conditions for the optimization problem (P1) read

$$\hat{\Sigma} - \hat{\Theta}^{-1} + \lambda\hat{Z} = 0, \tag{16}$$

where the matrix $\hat{Z}$ is the sub-differential of $\|.\|_{1,\text{off}}$ at the optimum $\hat{\Theta}$. Multiplying (16) by $\hat{\Theta}$ from both sides (which is equivalent to multiplying the vectorized equation (16) by $\hat{\Theta} \otimes \hat{\Theta}$), we obtain

$$\hat{\Theta}\hat{\Sigma}\hat{\Theta} - \hat{\Theta} + \hat{\Theta}\lambda\hat{Z}\hat{\Theta} = 0.$$

Adding $\hat{\Theta} - \Theta^*$ to both sides and rearranging gives

$$\underbrace{\hat{\Theta} + \hat{\Theta}\lambda\hat{Z}\hat{\Theta}}_{\hat{T}} - \Theta^* = -\Theta^*(\hat{\Sigma} - \Sigma^*)\Theta^* + \text{rem}, \tag{17}$$

where, denoting $W := \hat{\Sigma} - \Sigma^*$, we have

$$\text{rem} := -(\hat{\Theta} - \Theta^*)W\Theta^* - (\hat{\Theta}\hat{\Sigma} - I)(\hat{\Theta} - \Theta^*). \tag{18}$$

Bound I

$$\begin{aligned}\|\text{rem}\|_\infty &\leq \|(\hat{\Theta} - \Theta^*)W\Theta^*\|_\infty + \|(\hat{\Theta}\hat{\Sigma} - I)(\hat{\Theta} - \Theta^*)\|_\infty \\ &\leq \left|\left|\left|\hat{\Theta} - \Theta^*\right|\right|\right|_\infty \|W\Theta^*\|_\infty + \|\hat{\Theta}\hat{\Sigma} - I\|_\infty \left|\left|\left|\hat{\Theta} - \Theta^*\right|\right|\right|_\infty\end{aligned}$$

We can bound

$$\begin{aligned}\|\hat{\Sigma}\hat{\Theta} - I\|_\infty &= \|(\hat{\Sigma} - \Sigma^*)(\hat{\Theta} - \Theta^*) + \Sigma^*(\hat{\Theta} - \Theta^*) + (\hat{\Sigma} - \Sigma^*)\Theta^*\|_\infty \\ &\leq \|\hat{\Sigma} - \Sigma^*\|_\infty \left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_\infty + |\!|\!|\Sigma^*|\!|\!|_\infty \|\hat{\Theta} - \Theta^*\|_\infty + \|W\Theta^*\|_\infty\end{aligned}$$

Hence

$$\begin{aligned}\|\text{rem}\|_\infty \leq\ & \underbrace{\left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_\infty \|W\Theta^*\|_\infty}_{\text{rem}_1} + \underbrace{\|\hat{\Sigma} - \Sigma^*\|_\infty \left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_\infty^2}_{\text{rem}_2} \\ & + \underbrace{|\!|\!|\Sigma^*|\!|\!|_\infty \|\hat{\Theta} - \Theta^*\|_\infty \left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_\infty}_{\text{rem}_3} + \underbrace{\|W\Theta^*\|_\infty \left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_\infty}_{\text{rem}_1}\end{aligned}$$

In what follows, condition on the event $\mathcal{T}_n = \{\|\hat{\Sigma} - \Sigma^*\|_\infty \leq \delta_n\}$. Note that $|\!|\!|\Theta^*|\!|\!|_\infty = \max_i \|\Theta_i\|_1 \leq \max_i \|\Theta_i\|_2 \sqrt{d} \leq \Lambda_{\max}(\Theta^*)\sqrt{d}$.

By Lemma 9, part (a), on $\mathcal{T}$ it holds that $\hat{\Theta}_{S^c} = \Theta^*_{S^c}$. Thus $\hat{\Theta}$ has at most d nonzero entries per row. Hence it follows

$$|\!|\!|\Delta|\!|\!|_\infty = \left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_\infty \leq d\|\hat{\Theta} - \Theta^*\|_\infty. \tag{19}$$

Next by Lemma 9, part (b), we have the bound

$$\|\hat{\Theta} - \Theta^*\|_\infty \leq 2(1 + 8/\alpha)\kappa_{\Gamma^*}\delta_n.$$

We obtain

$$\begin{aligned}\|\text{rem}_1\|_\infty &\leq \|\Theta^* W\|_\infty |\!|\!|\Delta|\!|\!|_1 \leq 2(1 + 8/\alpha)\kappa_{\Gamma^*} d\delta_n \|\Theta^* W\|_\infty \\ \|\text{rem}_2\|_\infty &\leq |\!|\!|\Delta|\!|\!|_1^2 \|W\|_\infty \leq 4(1 + 8/\alpha)^2 \kappa_{\Gamma^*}^2 d^2 \delta_n^3 \\ \|\text{rem}_3\|_\infty &\leq \|\Delta\|_\infty |\!|\!|\Delta|\!|\!|_1 |\!|\!|\Sigma^*|\!|\!|_1 \leq 4(1 + 8/\alpha)^2 \kappa_{\Gamma^*}^2 \kappa_{\Sigma^*} d\delta_n^2.\end{aligned}$$

Hence conditioned on $\mathcal{T}$,

$$\begin{aligned}\|\text{rem}\|_\infty &\leq 4\max\{2(1 + 8/\alpha)\kappa_{\Gamma^*} d\delta_n \|\Theta^* W\|_\infty, \\ &\quad 4(1 + 8/\alpha)^2 \kappa_{\Gamma^*}^2 d^2 \delta_n^3, \\ &\quad 4(1 + 8/\alpha)^2 \kappa_{\Gamma^*}^2 \kappa_{\Sigma^*} d\delta_n^2\} \\ &= \mathcal{O}\left(\frac{1}{\alpha}\kappa_{\Gamma^*} \max\{d\delta_n \|\Theta^* W\|_\infty, \frac{1}{\alpha}\kappa_{\Gamma^*} d^2 \delta_n^3, \frac{1}{\alpha}\kappa_{\Gamma^*}\kappa_{\Sigma^*} d\delta_n^2\}\right).\end{aligned}$$

Bound II

Observe that

$$\left|\!\left|\!\left|\hat{\Theta}\right|\!\right|\!\right|_1 \leq \left|\!\left|\!\left|\hat{\Theta} - \Theta^*\right|\!\right|\!\right|_1 + |\!|\!|\Theta^*|\!|\!|_1 \leq 2(1 + 8/\alpha)\kappa_{\Gamma^*} d\delta_n + \sqrt{d}\Lambda_{\max}(\Theta^*).$$

By (18) and using the KKT conditions we have

$$\|\text{rem}\|_\infty = \|\Delta W\Theta^* - \hat{\Theta}(\hat{\Sigma} - \hat{\Theta}^{-1})\|_\infty$$

$$
\begin{aligned}
&\leq \|\Delta W\Theta^*\|_\infty + \left|\!\left|\!\left|\hat{\Theta}\right|\!\right|\!\right|_1 \|\lambda\hat{Z}\|_\infty \|\Delta\|_\infty \\
&\leq 2(1+8/\alpha)\kappa_{\Gamma^*} d\delta_n \|\Theta^* W\|_\infty \\
&\quad + 16\frac{8}{\alpha}4(1+8/\alpha)^2\kappa_{\Gamma^*}^2 d^2\delta_n^3 + 2\frac{8}{\alpha}(1+8/\alpha)\Lambda_{\max}(\Theta^*)\kappa_{\Gamma^*} d^{3/2}\delta_n^2 \\
&= \mathcal{O}\left(\frac{1}{\alpha}\kappa_{\Gamma^*}\max\{d\delta_n\|\Theta^* W\|_\infty, \frac{1}{\alpha^2}\kappa_{\Gamma^*} d^2\delta_n^3, \frac{1}{\alpha}\Lambda_{\max}(\Theta^*) d^{3/2}\delta_n^2\}\right). \square
\end{aligned}
$$

Proof of Lemma 1. (i) Under (C1), using Lemma 8 and by assumption (A2) we have

$$
\begin{aligned}
\|\Theta^* W\|_\infty &\leq |\!|\!|\Theta^*|\!|\!|_\infty \|W\|_\infty = \max_{i=1,\ldots,p} \|\Theta_i^*\|_1 \|W\|_\infty \\
&\leq \sqrt{d}\|\Theta_i^*\|_2 \|W\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{d\log p/n}).
\end{aligned}
$$

Then by Lemma 3, bound II,

$$
\begin{aligned}
\|\mathrm{rem}\|_\infty &= \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\alpha}\kappa_{\Gamma^*}\max\{d\delta_n\|\Theta^* W\|_\infty, \frac{1}{\alpha^2}\kappa_{\Gamma^*} d^2\delta_n^3, \frac{1}{\alpha}\Lambda_{\max}(\Theta^*) d^{3/2}\delta_n^2\}\right) \\
&= \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\alpha^2}\kappa_{\Gamma^*}\max\{d^{3/2}\log p/n, \frac{1}{\alpha}\kappa_{\Gamma^*} d^2(\log p/n)^{3/2}\}\right).
\end{aligned}
$$

(ii) Under (C2) and (A2), by Lemma 7 we have

$$
\|\Theta^* W\|_\infty = \max_{i,j=1,\ldots,p} |\Theta_i^* W e_j| = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}).
$$

Hence using Lemma 3, bound I and Lemma 8 we have

$$
\begin{aligned}
\|\mathrm{rem}\|_\infty &= \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\alpha}\kappa_{\Gamma^*}\max\{d\log p/n, \frac{1}{\alpha}\kappa_{\Gamma^*} d^2(\log p/n)^{3/2}, \frac{\kappa_{\Gamma^*}\kappa_{\Sigma^*}}{\alpha} d\log p/n\}\right) \\
&\overset{(a)}{=} \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\alpha^2}\kappa_{\Gamma^*}^2\kappa_{\Sigma^*} d\log p/n\right).
\end{aligned}
$$

Step (a) above follows by the sparsity assumption (13) and since $\kappa_{\Gamma^*} \gtrsim 1$. The latter statement follows by (A2), using that $\Lambda_{\min}((\Gamma^*_{SS})^{-1}) = 1/\Lambda_{\max}(\Gamma^*_{SS}) \geq 1/\Lambda_{\max}(\Gamma^*)$ and that the eigenvalues of Kronecker product satisfy $\Lambda_{\max}(\Sigma\otimes\Sigma) = \Lambda^2_{\max}(\Sigma)$. $\square$

Proof of Theorem 1. By Lemma 1, under the sub-Gaussianity condition (C1) or (C2), under the appropriate sparsity assumptions (combining (13) and (14), (15)) it holds that $\|\mathrm{rem}\|_\infty = o_{\mathbb{P}}(1)$. Hence for every (i,j) it holds that

$$
\sqrt{n}(\hat{T}_{ij} - \Theta^*_{ij}) = \frac{1}{\sqrt{n}}\sum_{k=1}^n (\Theta_i^{*T} X_k \Theta_j^{*T} X_k - \Theta^*_{ij}) + o_{\mathbb{P}}(1). \tag{20}
$$

It remains to prove that the scaled summation term in (20) weakly converges to the normal distribution. To this end, define

$$
Z_{ij,k} := \Theta_i^{*T} X_k \Theta_j^{*T} X_k - \Theta^*_{ij}.
$$

For each n fixed, $Z_{ij,1},\ldots,Z_{ij,n}$ are identically distributed r.v.s with mean $\mathbb{E}(Z_{ij,k}) = (\Theta_i^*)^T\Sigma^*\Theta_j^* - \Theta_{ij} = e_i^T\Theta^*\Sigma^*\Theta^*e_j - \Theta_{ij}^* = 0$.

Since each X_k has sub-Gaussian elements, the variance $\sigma_{ij}^2 = \mathrm{Var}(Z_{ij,k}) = \mathrm{Var}(\Theta_i^{*T}X_k\Theta_j^{*T}X_k)$ is finite. Denote $S_n = \sum_{k=1}^n Z_{ij,k}$. Then $s_n := \mathrm{var}(S_n) = n\sigma_{ij}^2$. Dividing (20) by $\sigma_{ij} > 0$, we obtain

$$\sqrt{n}(\hat{T}_{ij} - \Theta_{ij}^*)/\sigma_{ij} = S_n/s_n + o_{\mathbb{P}}(1)/\sigma_{ij}.$$

First note that by the assumption $1/\sigma_{ij} = \mathcal{O}(1)$ we have $o_{\mathbb{P}}(1)/\sigma_{ij} = o_{\mathbb{P}}(1)$.

(i) Sub-Gaussian design (C1)

To show $S_n/s_n \rightsquigarrow \mathcal{N}(0,1)$, we check the Lindeberg condition, i.e. for all $\varepsilon > 0$

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{k=1}^n \mathbb{E}(Z_{ij,k}^2\mathbf{1}(|Z_{ij,k}| > \varepsilon s_n)) = 0.$$

Observe that since $\{Z_{ij,k}\}_{k=1}^n$ are identically distributed for n fixed and $s_n^2 = n\sigma_{ij}^2$,

$$\sum_{k=1}^n \frac{1}{s_n^2}\mathbb{E}(Z_{ij,k}^2\mathbf{1}(|Z_{ij,k}| > \varepsilon s_n)) = \frac{1}{\sigma_{ij}^2}\mathbb{E}(Z_{ij,1}^2\mathbf{1}(|Z_{ij,1}| > \varepsilon s_n)).$$

Consequently, it remains to show that $\lim_{n\to\infty}\mathbb{E}(Z_{ij,1}^2\mathbf{1}(|Z_{ij,1}| > \varepsilon s_n)) = 0$. For any $c > 0$ we may rewrite

$$c = \int_0^\infty \mathbf{1}_{c>t}dt.$$

Therefore, applying the last observation and by Fubini's theorem we obtain

$$\int_\Omega |X|^2\mathbf{1}_{|X|>a}d\mathbb{P} = a^2\mathbb{P}(|X| > a) + 2\int_a^\infty u\mathbb{P}(|X| > u)du.$$

Then it follows

$$\begin{aligned}\mathbb{E}(Z_{ij,1}^2\mathbf{1}(|Z_{ij,1}| > \varepsilon\sigma_{ij}\sqrt{n})) \quad &\le \quad \varepsilon^2\sigma_{ij}^2 n\mathbb{P}(|Z_{ij,1}| > \varepsilon\sigma_{ij}\sqrt{n}) \\ &\quad + \int_{\varepsilon\sigma_{ij}\sqrt{n}}^\infty x\mathbb{P}(|Z_{ij,1}| > x)dx.\end{aligned}$$

To show that the limit of the right-hand side of the last inequality is 0 for $n\to\infty$, we use Lemma 4, by which it follows that $Z_{ij,1}$ satisfies a tail bound $\mathbb{P}(|Z_{ij,1}| > t) \le 4de^{-\frac{t}{c_1 d}}$. For a fixed $\varepsilon > 0$, putting $t := \varepsilon\sigma_{ij}\sqrt{n}$, we obtain

$$\mathbb{P}(|Z_{ij,1}| > \varepsilon\sigma_{ij}\sqrt{n}) \le 4de^{-\frac{\varepsilon\sigma_{ij}\sqrt{n}}{c_1 d}}.$$

Consequently, for the first term in (21) we have

$$\lim_{n\to\infty}\sigma_{ij}^2 nP(|Z_{ij,1}| > \varepsilon\sigma_{ij}\sqrt{n}) \le \lim_{n\to\infty}\sigma_{ij}^2 nde^{-\frac{\varepsilon\sigma_{ij}\sqrt{n}}{c_1 d}} = 0, \tag{21}$$

which follows by the sparsity assumption (8) that implies $d^{\frac{3}{2}} = o(\sqrt{n}/\log p)$. Next considering the limit of the last term in (21), we have

$$\int_{\varepsilon\sigma_{ij}\sqrt{n}}^{\infty} xP(|Z_{ij,1}| > x)dx \leq \int_{\varepsilon\sigma_{ij}\sqrt{n}}^{\infty} 4xde^{-\frac{x}{c_1 d}}dx.$$

In the integral, substitute $t := \frac{x}{\sigma_{ij}\sqrt{n}}$ to obtain

$$\lim_{n\to\infty} \int_{\sigma_{ij}\sqrt{n}}^{\infty} dxe^{-\frac{x}{d}}dx = \lim_{n\to\infty} \int_{1}^{\infty} d\sigma_{ij}^2 nte^{-\frac{\sigma_{ij}\sqrt{n}}{d}t}dt.$$

Again by the restriction on d and the Lebesgue dominated convergence it then follows that the limit of the integral is 0. In conclusion, we get $S_n/s_n \rightsquigarrow \mathcal{N}(0,1)$ for $n \to \infty$.

(ii) Sub-Gaussian design (C2)

Under (C2), we have a bound $\mathbb{P}(|Z_{ij,1}| > t) \lesssim e^{-t/(c_2K^2)}$ hence similarly as in (i), asymptotic normality follows. □

Lemma 4. *Let Θ^* satisfy assumption (A2) and let the random vector $X \in \mathbb{R}^p$ satisfy the sub-Gaussianity condition (C1) with $K = \mathcal{O}(1)$. Then for $t > c_0$ the random variable $Z := \Theta_i^{*T} X \Theta_j^{*T} X - \Theta_{ij}^*$ satisfies the following bound*

$$\mathbb{P}(|Z| > t) \leq 4de^{-\frac{t}{c_1 d}},$$

where c_0, c_1 do not depend on n.

Proof. Since $|\Theta_i^{*T} X| \leq \|\Theta_i\|_1 \max_{k:\Theta_{ki}^*\neq 0} |X^k| \leq \|\Theta_i^*\|_2\sqrt{d}\max_{k:\Theta_{ki}^*\neq 0}|X^k|$, and using the union bound and sub-Gaussianity of $X_n^k/\sqrt{\Sigma_{kk}^*}$ by assumption (C1) gives

$$\begin{aligned}
\mathbb{P}(|\Theta_i^{*T} X| > t) &\leq \mathbb{P}\left(\max_{k:\Theta_{ki}^*\neq 0} |X^k| > \frac{t}{\sqrt{d}\|\Theta_i^*\|_2}\right) \\
&\leq d \max_{k:\Theta_{ki}^*\neq 0} \mathbb{P}\left(|X^k|/\sqrt{\Sigma_{kk}^*} > \frac{t}{\sqrt{d}\|\Theta_i^*\|_2\sqrt{\Sigma_{kk}^*}}\right) \\
&\leq 2d\exp\left(-\frac{t^2}{6K^2 d\|\Theta_i^*\|_2^2 \max_k \Sigma_{kk}^*}\right).
\end{aligned}$$

Then for $\Theta_i^{*T} X \Theta_j^{*T} X$ we obtain the bound

$$\mathbb{P}(|\Theta_i^{*T} X \Theta_j^{*T} X| > t) \leq 4d\exp\left(-\frac{t}{6K^2 d\Lambda_{\max}^2(\Theta^*)\max_k \Sigma_{kk}^*}\right).$$

Under (A2), Σ_{kk}^* and $\|\Theta_i^*\|_2 \leq \Lambda_{\max}(\Theta^*)$ are uniformly bounded in n. Thus for $Z = \Theta_i^{*T} X \Theta_j^{*T} X - \Theta_{ij}^*$, there exist constants c_0, c_1, c_2 not depending on n such that for $t > c_0 > |\Theta_{ij}^*|$

$$\mathbb{P}(|Z| > t) \leq 4de^{-\frac{t-|\Theta_{ij}^*|}{c_2 d}} \leq 4de^{-\frac{t}{c_1 d}},$$

since $|\Theta_{ij}^*|/d$ is bounded by (A2) and $d \geq 1$. □

Proof of Lemma 2. Since $X \sim \mathcal{N}(0, \Sigma^*)$, then $\Theta^* X \sim \mathcal{N}(0, \Theta^*)$, hence

$$\sigma_{ij}^2 = \mathrm{Var}(\Theta_i^{*T} X \Theta_j^{*T} X) = \mathrm{Var}(e_i^T \Theta^{*T} X X^T \Theta^* e_j) = \mathrm{Var}(e_i^T Z Z^T e_j),$$

where $Z \sim \mathcal{N}(0, \Theta^*)$. Thus

$$\begin{aligned} \sigma_{ij}^2 &= \mathrm{Var}(Z^i Z^j) = \mathbb{E}((Z^i)^2 (Z^j)^2) - \mathbb{E}(Z^i Z^j)^2 \\ &= \Theta_{ii}^* \Theta_{jj}^* + 2\Theta_{ij}^{*\,2} - \Theta_{ij}^{*\,2} = \Theta_{ii}^* \Theta_{jj}^* + \Theta_{ij}^{*\,2}. \end{aligned}$$

By assumption (A2), $\Theta_{ii}^* \Theta_{jj}^* + \Theta^{*2}_{ij} \geq \Lambda_{\min}^2(\Theta^*) \geq \frac{1}{L^2} > 0$, where $L \asymp 1$, therefore $1/\sigma_{ij} = \mathcal{O}(1)$.

By Lemma 9 and Lemma 8 we have

$$\begin{aligned} |\hat{\sigma}_{ij}^2 - \sigma_{ij}^2| &\leq |\hat{\Theta}_{ii}\hat{\Theta}_{jj} - \Theta_{ii}^*\Theta_{jj}^*| + |\hat{\Theta}_{ij}^2 - \Theta^{*2}_{ij}| \\ &\leq |\Delta_{ii}\Delta_{jj} + \Theta_{ii}^*\Delta_{jj} + \Theta_{jj}^*\Delta_{ii}| \\ &\quad + |\Delta_{ij}(\Delta_{ij} + 2\Theta_{ij}^*)| \\ &= \mathcal{O}_{\mathbb{P}}\left(1/\alpha\kappa_{\Gamma^*}\sqrt{\frac{\log p}{n}}\right), \end{aligned} \tag{22}$$

where we used assumption (A2) and the sparsity assumption (13). □

Concentration for sub-Gaussian design (C2)

Lemma 5. *Let $\alpha, \beta \in \mathbb{R}^p$ such that $\|\alpha\|_2 \leq M, \|\beta\|_2 \leq M$. Let $X_k \in \mathbb{R}^p$ satisfy the sub-Gaussianity assumption (C2) with a constant $K > 0$. Then for $m \geq 2$,*

$$\mathbb{E}|\alpha^T X_k X_k^T \beta - \mathbb{E}\alpha^T X_k X_k^T \beta|^m / (2M^2K^2)^m \leq \frac{m!}{2}.$$

Consequently, we may apply Bernstein inequality (Lemma 14.9 in [5]).

Lemma 6. *Let $\alpha, \beta \in \mathbb{R}^p$ such that $\|\alpha\|_2 \leq M, \|\beta\|_2 \leq M$. Let $X_k \in \mathbb{R}^p$ satisfy the sub-Gaussianity assumption (C2) with a constant $K > 0$. For all $t > 0$*

$$\mathbb{P}\left(|\alpha^T \hat{\Sigma} \beta - \alpha^T \Sigma \beta| / (2M^2K^2) > t + \sqrt{2t}\right) \leq 2e^{-nt}.$$

Lemma 7. *Assume $\|\alpha_i\|_2 \leq M, \|\beta\|_2 \leq M$ for all $i = 1, \ldots, p$ and (C2) with K. For all $t > 0$ it holds*

$$\mathbb{P}\left(\max_{i=1,\ldots,p} |\alpha_i^T (\hat{\Sigma} - \Sigma)\beta| / (2M^2K^2) > t + \sqrt{2t} + \sqrt{\frac{2\log(2p)}{n}} + \frac{\log(2p)}{n}\right) \leq e^{-nt}.$$

Proof. By Lemma 5 we have

$$\mathbb{E}|\alpha^T X_k X_k^T \beta - \mathbb{E}\alpha^T X_k X_k^T \beta|^m / (2M^2K^2)^m \leq \frac{m!}{2}.$$

Lemma 14.13 in [5] gives the claim. □

Proof of Lemma 5. By assumption we have $\|\alpha\|_2 \leq M, \|\beta\|_2 \leq M$. Consequently, by the sub-Gaussianity assumption with a constant K we obtain

$$\mathbb{E}e^{|X_k^T\alpha|^2/(MK)^2} \leq 2$$

and likewise

$$\mathbb{E}e^{|X_k^T\beta|^2/(MK)^2} \leq 2.$$

Next by the inequality $ab \leq a^2/2 + b^2/2$ (for any $a, b \in \mathbb{R}$) and by the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}\mathbb{E}e^{|\alpha^T X_k X_k^T \beta|/(MK)^2} &\leq \mathbb{E}e^{|X_k^T\alpha|^2/(MK)^2/2} e^{|X_k^T\beta|^2/(MK)^2/2} \\ &\leq \{\mathbb{E}e^{|X_k^T\alpha|^2/(MK)^2}\}^{1/2}\{\mathbb{E}e^{|X_k^T\beta|^2/(MK)^2}\}^{1/2} \\ &\leq 2.\end{aligned}$$

By the Taylor expansion, we have the inequality

$$1 + \frac{1}{m!}\mathbb{E}|\alpha^T X_k X_k^T \beta|^m/(MK)^{2m} \leq \mathbb{E}e^{|\alpha^T X_k X_k^T \beta|/(MK)^2}$$

Next it follows

$$\begin{aligned}\mathbb{E}|\alpha^T X_k X_k^T \beta - \mathbb{E}\alpha^T X_k X_k^T \beta|^m/(MK)^{2m} &\leq 2^{m-1}\mathbb{E}|\alpha^T X_k X_k^T \beta|^m/(MK)^{2m} \\ &\leq 2^{m-1} m!(\mathbb{E}e^{|\alpha^T X_k X_k^T \beta|/(MK)^2} - 1) \\ &= 2^{m-1} m! = \frac{m!}{2}2^m.\end{aligned}$$

And thus

$$\mathbb{E}|\alpha^T X_k X_k^T \beta - \mathbb{E}\alpha^T X_k X_k^T \beta|^m/(2M^2K^2)^m \leq \frac{m!}{2}.$$ □

Appendix A: Tail bounds and rates of convergence

The following Lemma from [26] gives probabilistic bounds for the event $\mathcal{T}_n$ if X satisfies (C1).

Lemma 8 ([26], sub-Gaussian model)**.** *Let $X = (X^1, \ldots, X^p)$ be independent, distributed as X with $\mathbb{E}X = 0$, $cov(X) = \Sigma^* = (\Theta^*)^{-1}$ and satisfying (C1) with $K > 0$. Then for*

$$\delta_\tau(n, r) = 8(1 + 12K^2)\max_i \Sigma^*_{ii}\sqrt{2\frac{\log(4r)}{n}},$$

and for every $\gamma > 2$ and each n such that $\delta_\tau(n, p^\gamma) < 8(1 + 12K^2)\max_i \Sigma^*_{ii}$ we have

$$\mathbb{P}\left(\|\hat{\Sigma} - \Sigma^*\|_\infty \geq \delta_\tau(n, p^\gamma)\right) \leq \frac{1}{p^{\gamma-2}}.$$

We restate a result on rates of convergence of the graphical Lasso from [26] in Lemma 9.

Lemma 9 (Theorem 1, [26])**.** *Suppose that $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent and distributed as $X = (X^1, \ldots, X^p)$ with $\mathbb{E}X = 0$, $cov(X) = \Sigma^*$. Let $\Theta^* = (\Sigma^*)^{-1}$ exist and satisfy the irrepresentability condition (A1) with a constant $\alpha \in (0, 1]$. Denote S^c to be the set of indices that correspond to zero entries of Θ^*. Let $\hat{\Theta}_n$ be the solution to the optimization problem (P1) with tuning parameter $\lambda_n = \frac{8}{\alpha}\delta_n$ for some $\delta_n > 0$ and suppose that the sparsity assumption*

$$d \leq \frac{1}{6(1+8/\alpha)\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa_{\Sigma^*}^3\kappa_{\Gamma^*}^2\}\delta_n}$$

is satisfied. Then on $\mathcal{T}_n = \{\|\hat{\Sigma} - \Sigma\|_\infty \leq \delta_n\}$ it holds

(a) $\hat{\Theta}_{S^c} = \Theta^*_{S^c}$
(b) $\|\hat{\Theta} - \Theta^*\|_\infty \leq 2(1 + 8/\alpha)\kappa_{\Gamma^*}\delta_n$.

For instance, if the observations $X_1, \ldots, X_n$ are sub-Gaussian (C1) with $K = \mathcal{O}(1)$, $\max_i \Sigma^*_{ii} = \mathcal{O}(1)$ then Lemma 8 gives $\|\hat{\Sigma} - \Sigma^*\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n})$.

Hence when the observations are sub-Gaussian, Lemma 9 implies the following convergence rates for the graphical Lasso estimator. If the quantities $\kappa_{\Sigma^*} = \mathcal{O}(1)$, $\kappa_{\Gamma^*} = \mathcal{O}(1)$, $1/\alpha = \mathcal{O}(1)$, the sparsity assumption $d = o(\sqrt{n/\log p})$ is satisfied (equivalently $n = \Omega(d^2 \log p)$), then for suitably chosen $\lambda_n \asymp \sqrt{\log p/n}$ we have

$$\|\hat{\Theta} - \Theta^*\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}).$$

References

[1] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[2] Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies* **81** 608–650. MR3207983

[3] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics* **41** 802–837. MR3099122

[4] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604. MR2485008

[5] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer. MR2807761

[6] Cai, T., Liu, W. and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607. MR2847973

[7] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351. MR2382644

[8] Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association* **106** 608–625. MR2847974

[9] Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive Lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics* **41**. MR3113809
[10] d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568
[11] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
[12] Greene, W. H. (2011). *Econometric Analysis*. Prentice Hall.
[13] Javanmard, A. and Montanari, A. (2013a). Confidence intervals and hypothesis testing for high-dimensional regression. *ArXiv:1306.3171*. MR3277152
[14] Javanmard, A. and Montanari, A. (2013b). Model selection for high-dimensional regression under the generalized irrepresentability condition. In *Advances in Neural Information Processing Systems 26* (C. j. c. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. q. Weinberger, eds.) 3012–3020.
[15] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* **29** 295–327. MR1863961
[16] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* **28** 1356–1378. MR1805787
[17] Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford. MR1419991
[18] Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856
[19] Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* **34** 2554–2591. MR2291510
[20] Meinshausen, N. (2013). Assumption-free confidence intervals for groups of variables in sparse high-dimensional regression. *ArXiv:1309.3489*. MR3066380
[21] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462. MR2278363
[22] Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104** 1671–1681. MR2750584
[23] Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557. MR3025133
[24] Ng, B., G. Varoquaux, J.-B. P. and Thirion, B. (2013). A novel sparse group gaussian graphical model for functional connectivity estimation. *Information Processing in Medical Imaging*.

[25] Nickl, R. and van de Geer, S. (2012). Confidence sets in sparse regression. *Annals of Statistics* **41** 2852–2876. MR3161450

[26] Ravikumar, P., Raskutti, G., Wainwright, M. J. and Yu, B. (2008). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980. MR2836766

[27] Ren, Z., Sun, T., Zhang, C. H. and Zhou, H. H. (2013). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *ArXiv:1309.6024*. MR3346695

[28] Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515. MR2417391

[29] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**. MR2183942

[30] Städler, N. and Mukherjee, S. (2013). Two-sample testing in high-dimensional models. *Annals of Applied Statistics* **7** 1837–2457.

[31] Sun, T. and Zhang, C. H. (2012). Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research* **14** 3385–3418. MR3144466

[32] van de Geer, S. (2014a). Statistical theory for high-dimensional models. *ArXiv:1309.3489*.

[33] van de Geer, S. (2014b). Worst possible sub-directions in high-dimensional models. *ArXiv:1403.7023*. MR3181133

[34] van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3** 1360–1392. MR2576316

[35] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* **42** 1166–1202. MR3224285

[36] Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics* **37** 2178. MR2543689

[37] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* **11** 2261–2286. MR2719856

[38] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 1–17. MR2367824

[39] Zhang, C. H. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B* **76** 217–242. MR3153940

[40] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563. MR2274449