Honest confidence regions and optimality in high-dimensional precision matrix estimation

Journal Article

Author(s):
Janková, Jana; van de Geer, Sara

Publication date:
2017-03

Permanent link:
https://doi.org/10.3929/ethz-b-000206094

Originally published in:
TEST 26(1), https://doi.org/10.1007/s11749-016-0503-5

TEST (2017) 26:143–162
DOI 10.1007/s11749-016-0503-5

ORIGINAL PAPER

Honest confidence regions and optimality in high-dimensional precision matrix estimation

Jana Janková[1] · **Sara van de Geer**[1]

Received: 20 July 2015 / Accepted: 4 September 2016 / Published online: 14 September 2016

Abstract We propose methodology for estimation of sparse precision matrices and statistical inference for their low-dimensional parameters in a high-dimensional setting where the number of parameters p can be much larger than the sample size. We show that the novel estimator achieves minimax rates in supremum norm and the low-dimensional components of the estimator have a Gaussian limiting distribution. These results hold uniformly over the class of precision matrices with row sparsity of small order $\sqrt{n}/\log p$ and spectrum uniformly bounded, under a sub-Gaussian tail assumption on the margins of the true underlying distribution. Consequently, our results lead to uniformly valid confidence regions for low-dimensional parameters of the precision matrix. Thresholding the estimator leads to variable selection without imposing irrepresentability conditions. The performance of the method is demonstrated in a simulation study and on real data.

Keywords Precision matrix · Sparsity · Inference · Asymptotic normality · Confidence regions

Mathematics Subject Classification 62J07 · 62F12

Electronic supplementary material The online version of this article (doi:10.1007/s11749-016-0503-5) contains supplementary material, which is available to authorized users.

✉ Jana Janková
jankova@stat.math.ethz.ch

Sara van de Geer
geer@stat.math.ethz.ch

[1] Seminar for Statistics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland

1 Introduction

We consider the problem of estimation of the inverse covariance matrix in a high-dimensional setting, where the number of parameters p can significantly exceed the sample size n. Suppose that we are given an $n \times p$ design matrix $\mathbf{X}$, where the rows of $\mathbf{X}$ are p-dimensional i.i.d. random vectors from an unknown distribution with mean zero and covariance matrix $\Sigma_0 \in \mathbb{R}^{p \times p}$. We denote the precision matrix by $\Theta_0 := \Sigma_0^{-1}$, assuming that the inverse of Σ_0 exists.

The problem of estimating the precision matrix arises in a wide range of applications. Precision matrix estimation in particular plays an important role in graphical models that have become a popular tool for representing dependencies within large sets of variables. Suppose that we associate the variables $X_1, \ldots, X_p$ with the vertex set $\mathcal{V} = \{1, \ldots, p\}$ of an undirected graph $G = (\mathcal{V}, \mathcal{E})$ with an edge set $\mathcal{E}$. A graphical model G represents the conditional dependence relationships between the variables, namely, every pair of variables not contained in the edge set is conditionally independent given all remaining variables. If the vector $(X_1, \ldots, X_p)$ is normally distributed, each edge corresponds to a non-zero entry in the precision matrix (Lauritzen 1996). Practical examples of applications of graphical modeling include modeling of brain connectivity based on FMRI brain analysis (Ng et al. 2013), genetic networks, financial data processing, social network analysis, and climate data analysis.

A lot of work has been done on methodology for *point* estimation of precision matrices. We discuss some of the approaches below, but a selected list of papers includes, for instance, Meinshausen and Bühlmann (2006), Friedman et al. (2008), Bickel and Levina (2008), Yuan (2010), Cai et al. (2011) and Sun and Zhang (2012). A common approach assumes that the precision matrix is sufficiently sparse and employs the ℓ_1-penalty to induce a sparse estimator. The main goal of these works is to show that under some regularity conditions, the sparse estimator behaves almost as the oracle estimator that has the knowledge of the true sparsity pattern.

Our primary interest in this paper lies not in point estimation, but we aim to quantify uncertainty of estimation by providing *interval* estimates for the entries of the precision matrix. The challenge of this problem arises, since asymptotics of regularized estimators which are the main tool in high-dimensional estimation is not easily tractable (Knight and Fu 2000), as opposed to the classical setting when the dimension of the unknown parameter is fixed.

1.1 Overview of related work

Methodology for inference in high-dimensional models has been mostly studied in the context of linear and generalized linear regression models. From the work on linear regression models, we mention the paper by Zhang and Zhang (2014) where a semi-parametric projection approach using the Lasso methodology (Tibshirani 1996) was proposed, which was further developed and studied in van de Geer et al. (2013). The approach leads to asymptotically normal estimation of the regression coefficients, and an extension of the method to generalized linear models is given in van de Geer et al. (2013). The method requires sparsity of small order $\sqrt{n}/\log p$ in the high-dimensional parameter vector and uses ℓ_1-norm error bound of the Lasso. Further alternative methods for inference in the linear model have been proposed and studied in Javanmard and Montanari (2014), Belloni et al. (2014), and bootstrapping approach was suggested in Chatterjee and Lahiri (2013), Chatterjee and Lahiri (2011).

Other lines of work on inference for high-dimensional models suggest post-model selection procedures, where in the first step, a regularized estimator is used for model selection, and in the second step, e.g., a maximum likelihood estimator is applied on the selected model. In the linear model, simple post-model selection methods have been proposed, e.g., in Javanmard and Montanari (2013), Candes and Tao (2007). These approaches are, however, only guaranteed to work under irrepresentability and beta-min conditions (see Bühlmann and van de Geer 2011). Especially in view of inference, beta-min conditions, which assume that the non-zero parameters are sufficiently large in absolute value, should be avoided.

In this paper, we consider estimation of precision matrices, which is a problem related to linear regression; however, it is a non-linear problem, and thus, it requires a more involved treatment. One approach to precision matrix estimation is based on regularization of the maximum likelihood in terms of the ℓ_1-penalty. This approach is typically referred to as the graphical Lasso, and has been studied in detail in several papers, see Friedman et al. (2008), Rothman et al. (2008), Ravikumar et al. (2008) and Yuan and Lin (2007). Another common approach to precision matrix estimation is based on projections. This approach reduces the problem to a series of regression problems and estimates each column of the precision matrix using a Lasso estimator or Dantzig selector (Candes and Tao 2007). The idea was first introduced in Meinshausen and Bühlmann (2006) as neighborhood selection for Gaussian graphical models and further studied in Yuan (2010), Cai et al. (2011) and Sun and Zhang (2012).

Methodology leading to statistical inference for the precision matrix has been studied only recently. The work Ren et al. (2015) proposes to use a more involved variation of the regression approach to obtain an estimator which leads to statistical inference. This approach leads to an estimator of the precision matrix which is elementwise asymptotically normal, under row sparsity of order $\sqrt{n}/\log p$, bounded spectrum of the true precision matrix and Gaussian distribution of the sample. The paper Janková and van de Geer (2015) proposes a method for statistical inference based on the graphical Lasso. The work introduces a desparsified estimator based on the graphical Lasso, which is also shown to be elementwise asymptotically normal.

1.2 Contributions and outline

We propose methodology leading to honest confidence intervals and testing for low-dimensional parameters of the precision matrix, without requiring irrepresentability conditions or beta-min conditions to hold. Our work is motivated by the semi-parametric approach in van de Geer et al. (2013) and is a follow-up of the work Janková and van de Geer (2015). Compared to the previous work on statistical inference for precision matrices, this methodology has several advantages. First, the estimator we propose is a simple modification of the nodewise Lasso estimator proposed in Meinshausen and Bühlmann (2006). Hence, the estimator is easy to implement, and efficient solutions are available on the computational side. Second, the novel estimator enjoys a range of optimality properties and leads to statistical inference under mild conditions. First, the asymptotic distribution of low-dimensional components of the estimator is shown to be Gaussian. This holds uniformly over the class of precision matrices with row sparsity of order $o(\sqrt{n}/\log p)$, spectrum uniformly bounded in n and sub-Gaussian margins of the underlying distribution. This results in honest confidence regions (Li 1989) for low-dimensional parameters. The proposed estimator achieves rate optimality, as shown in Sect. 3.2. Moreover, the desparsified estimator may be thresholded to guarantee variable selection without imposing irrepresentable conditions. The computational cost of the method is order $\mathcal{O}(p)$ Lasso regressions for estimation of all parameters and two Lasso regressions for a single parameter.

This paper is organized as follows. Section 2 introduces the methodology. Section 3 contains the main theoretical results for estimation and inference, and in Sect. 3.4 the suggested method is applied to variable selection. Section 4 provides a comparison with related work. Section 5 illustrates the theoretical results in a simulation study. In Sect. 6, we analyze two real data sets and apply our method to variable selection. Section 7 contains a brief summary of the results. Finally, the proofs were deferred to the Online Resource 1.

Notation For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $p \in (0, \infty]$, we use $\|x\|_p$ to denote the $p-$norm of x in the classical sense. We denote $\|x\|_0 = |\{i : x_i \neq 0\}|$. For a matrix $A \in \mathbb{R}^{d\times d}$, we use the notations $|||A|||_\infty = \max_i \|e_i^T A\|_1$, $|||A|||_1 = \left|\left|\left|A^T\right|\right|\right|_\infty$ and $\|A\|_\infty = \max_{i,j} |A_{ij}|$. The symbol $\text{vec}(A)$ denotes the vectorized version of a matrix A obtained by stacking the rows of A on each other. By e_i, we denote a p-dimensional vector of zeros with one at position i. For real sequences f_n, g_n, we write $f_n = O(g_n)$ if $|f_n| \le C|g_n|$ 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 $1/f_n = \mathcal{O}(1/g_n)$ hold. Finally, $f_n = o(g_n)$ if $\lim_{n\to\infty} f_n/g_n = 0$. Furthermore, for a sequence of random variables x_n, we write $x_n = \mathcal{O}_{\mathbb{P}}(1)$ if x_n is bounded in probability and we write $x_n = \mathcal{O}_{\mathbb{P}}(r_n)$ if $x_n/r_n = \mathcal{O}_{\mathbb{P}}(1)$. We write $x_n = o_{\mathbb{P}}(1)$ if x_n converges in probability to zero.

Let $\rightsquigarrow$ denote the convergence in distribution and $\xrightarrow{P}$ the convergence in probability. Let Φ denote the cumulative distribution function of a standard normal random variable. By $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$, we denote the minimum and maximum eigenvalue of A, respectively. Let $a \vee b$, $a \wedge b$ denote $\max(a, b)$, $\min(a, b)$, respectively. We use letters C, c to denotes universal constants. These are used in the proofs repeatedly to denote possibly different constants.

2 Desparsified nodewise Lasso

Our methodology is a simple modification of the nodewise Lasso estimator proposed in Meinshausen and Bühlmann (2006). The idea is to remove the bias term which arises in the nodewise Lasso estimator due to ℓ_1-penalty regularization. This approach in inspired by literature on semiparametric statistics Bickel et al. (1993), van der Vaart (2000). We note several papers have used this idea in the context of high-dimensional sparse estimation, see Zhang and Zhang (2014), van de Geer et al. (2013), van de Geer (2016), Javanmard and Montanari (2014), Janková and van de Geer (2015).

We first summarize the nodewise Lasso method introduced in Meinshausen and Bühlmann (2006) and discuss some of its properties. This method estimates an unknown precision matrix using the idea of projections to approximately invert the sample covariance matrix. For each $j = 1, \dots, p$, we define the vector $\gamma_j = \{\gamma_{j,k}, k \neq j\}$ as follows:

$$\gamma_j := \arg \min_{\gamma \in \mathbb{R}^{p-1}} \mathbb{E}\|X_j - \mathbf{X}_{-j}\gamma\|_2^2/n \tag{1}$$

and denote $\eta_j := X_j - \mathbf{X}_{-j}\gamma_j$ and the noise level by $\tau_j^2 = \mathbb{E}\eta_j^T\eta_j/n$. We define the column vector $\Gamma_j := (-\gamma_{j,1}, \dots, -\gamma_{j,j-1}, 1, -\gamma_{j,j+1}, \dots, -\gamma_{j,p})^T$. Then, one may show:

$$\Theta_0 = (\Theta_1^0, \dots, \Theta_p^0) = (\Gamma_1/\tau_1^2, \dots, \Gamma_p/\tau_p^2), \tag{2}$$

where Θ_j^0 is the jth column of Θ_0. Hence, the precision matrix Θ_0 may be recovered from the partial correlations $\gamma_{j,k}$ and from the noise level τ_j^2. In our problem, we are only given the design matrix $\mathbf{X}$. The idea of nodewise Lasso is to estimate the partial correlations and the noise levels by doing a projection of every column of the design matrix on all the remaining columns. In low-dimensional settings, this procedure would simply recover the sample covariance matrix $\mathbf{X}^T\mathbf{X}/n$. However, due to the high-dimensionality of our setting, the matrix $\mathbf{X}^T\mathbf{X}/n$ is not invertible and we can only do approximate projections. If we assume sparsity in the precision matrix (and thus also in the partial correlations), this idea can be effectively carried out using the Lasso. Hence, for each $j = 1, \dots, p$ define the estimators of the regression coefficients, $\hat{\gamma}_j = \{\hat{\gamma}_{j,k}, k = 1, \dots, p, j \neq k\} \in \mathbb{R}^{p-1}$, as follows:

$$\hat{\gamma}_j := \arg \min_{\gamma \in \mathbb{R}^{p-1}} \|X_j - \mathbf{X}_{-j}\gamma\|_2^2/n + 2\lambda_j\|\gamma\|_1. \tag{3}$$

We further define the column vectors

$$\hat{\Gamma}_j := (-\hat{\gamma}_{j,1}, \dots, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, \dots, -\hat{\gamma}_{j,p})^T,$$

and estimators of the noise level

$$\hat{\tau}_j^2 := \|X_j - \mathbf{X}_{-j}\hat{\gamma}_j\|_2^2/n + \lambda_j\|\hat{\gamma}_j\|_1,$$

for $j = 1, \dots, p$. Finally, we define the jth column of the nodewise Lasso estimator $\hat{\Theta}$ as:

$$\hat{\Theta}_j := \hat{\Gamma}_j/\hat{\tau}_j^2. \tag{4}$$

The estimator $\hat{\Theta}_j$ of the precision matrix was studied in several papers (following Meinshausen and Bühlmann 2006) and has been shown to enjoy oracle properties under mild conditions on the model. These conditions include bounded spectrum of the precision matrix, row sparsity of small order $n/\log p$ and a sub-Gaussian distribution of the rows of $\mathbf{X}$ (alternatively to sub-Gaussianity, one may assume that the covariates are bounded as in van de Geer et al. 2013). Our approach uses the nodewise Lasso estimator as an initial estimator. The next step involves debiasing or desparsifying, which may be viewed as one step using the Newton–Raphson scheme for numerical optimization. This is equivalent to "inverting" the Karush–Kuhn–Tucker (KKT) conditions by the inverse of the Fisher information as in van de Geer et al. (2013). The challenge then also comes from the need to estimate the Fisher information matrix which is a $p^2 \times p^2$ matrix. We show that the estimator $\hat{\Theta}$ can be used in a certain way to create a surrogate of the inverse Fisher information matrix. Since the estimator $\hat{\Theta}_j$ can be characterized by its KKT conditions, it is convenient to work with these, conditions to derive the new desparsified estimator. Consider, hence, the KKT conditions for the optimization problem (3):

$$-\mathbf{X}_{-j}^T(X_j - \mathbf{X}_{-j}\hat{\gamma}_j)/n + \lambda_j \hat{\kappa}_j = 0, \tag{5}$$

for $j = 1, \ldots, p$, where $\hat{\kappa}_j$ is the subdifferential of the function $\gamma_j \mapsto \|\gamma_j\|_1$ at $\hat{\gamma}_j$, i.e.,

$$\hat{\kappa}_{j,k} = \begin{cases} \operatorname{sign}(\hat{\gamma}_{j,k}) & \text{if } \hat{\gamma}_{j,k} \neq 0 \\ a_{j,k} \in [-1, 1] & \text{otherwise,} \end{cases}$$

where $k \in \{1, \ldots, p\}\backslash\{j\}$. If we define $\hat{Z}_j$ to be a $p \times 1$ vector

$$\hat{Z}_j := (\hat{\kappa}_{j,1}, \ldots, \hat{\kappa}_{j,j-1}, 0, \hat{\kappa}_{j,j+1}, \ldots, \hat{\kappa}_{j,p})/\hat{\tau}_j^2,$$

then the KKT conditions may be equivalently stated as follows:

$$\hat{\Sigma}\hat{\Theta}_j - e_j - \lambda_j \hat{Z}_j = 0, \text{ for } j = 1, \ldots, p, \tag{6}$$

where $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ is the sample covariance matrix. This is shown in Lemma 11 in the Online Resource 1. Consequently, the KKT conditions (6) imply a bound $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j/\hat{\tau}_j^2$ for each $j = 1, \ldots, p$, which will be useful later. Note that the KKT conditions may be equivalently summarized in a matrix form as $\hat{\Sigma}\hat{\Theta} - I - \hat{Z}\Lambda = 0$, where the columns of $\hat{Z}$ are given by $\hat{Z}_j$ for $j = 1, \ldots, p$ and Λ is a diagonal matrix with elements $(\lambda_1, \ldots, \lambda_p)$.

Multiplying the KKT conditions (6) by $\hat{\Theta}_i$, we obtain

$$\hat{\Theta}_i^T(\hat{\Sigma}\hat{\Theta}_j - e_j) - \hat{\Theta}_i^T \lambda_j \hat{Z}_j = 0.$$

Then, we note that adding $\hat{\Theta}_{ij} - \Theta_{ij}^0$ to both sides and rearranging, we get:

$$\begin{aligned} \hat{\Theta}_{ij} - \hat{\Theta}_i^T \lambda_j \hat{Z}_j - \Theta_{ij}^0 &= \hat{\Theta}_{ij} - \hat{\Theta}_i^T(\hat{\Sigma}\hat{\Theta}_j - e_j) - \Theta_{ij}^0 \\ &= -(\Theta_i^0)^T(\hat{\Sigma} - \Sigma_0)\Theta_j^0 + \tilde{\Delta}_{ij}, \end{aligned} \tag{7}$$

where $\tilde{\Delta}_{ij} = -(\hat{\Theta}_i - \Theta_i^0)^T(\hat{\Sigma}\hat{\Theta}_j - e_j) - (\hat{\Theta}_j - \Theta_j^0)^T(\hat{\Sigma}\Theta_i^0 - e_i)$ is a term which can be shown to be $o_{\mathbb{P}}(n^{-1/2})$ under certain conditions (Lemma 1). Hence, we define the desparsified nodewise Lasso estimator:

$$\hat{T} := \hat{\Theta} - \hat{\Theta}^T(\hat{\Sigma}\hat{\Theta} - I) = \hat{\Theta} + \hat{\Theta}^T - \hat{\Theta}^T\hat{\Sigma}\hat{\Theta}. \tag{8}$$

3 Theoretical results

In this section, we inspect the asymptotic behavior of the desparsified nodewise Lasso estimator (8). In particular, we consider the limiting distribution of individual entries of $\hat{T}$ and show the convergence to the Gaussian distribution is uniform over the considered model. For construction of confidence intervals, we consider estimators of the asymptotic variance of the proposed estimator, both for Gaussian and sub-Gaussian design. We derive convergence rates of the method in supremum norm and consider application to variable selection. For completeness, in Lemma 6 in the Online Resource 1, we summarize convergence rates of the nodewise Lasso estimator. The result is essentially the same as Theorem 2.4 in van de Geer et al. (2013), so the proof of the common parts is omitted. Recall that $\mathcal{V} = \{1, \ldots, p\}$ and we define the row sparsity by $s_j := \|\Theta_j^0\|_0$, maximum row sparsity by $s := \max_{1\le j\le p} s_j$ and the coordinates of non-zero entries of the precision matrix by $S_0 := \{(i, j) \in \mathcal{V} \times \mathcal{V} : \Theta_{ij}^0 \neq 0\}$. For the analysis below, we will need the following conditions.

A1 (Bounded spectrum) The inverse covariance matrix $\Theta_0 := \Sigma_0^{-1}$ exists and there exists a universal constant $L \geq 1$, such that

$$1/L \leq \Lambda_{\min}(\Theta_0) \leq \Lambda_{\max}(\Theta_0) \leq L.$$

A2 (Sparsity) $\frac{s \log p}{n} = o(1)$.

A3 (Sub-Gaussianity condition) Suppose that the design matrix $\mathbf{X}$ has uniformly sub-Gaussian rows X_i, i.e., there exists a universal constant K, such that

$$\sup_{\alpha\in\mathbb{R}^p:\|\alpha\|_2\le 1} \mathbb{E}\exp\left(|\alpha^T X_i|^2/K^2\right) \leq 2 \quad (i = 1, \ldots, n).$$

The lower bound in A1 guarantees that the noise level $\tau_j^2 = 1/\Theta_{jj}^0$ does not diverge. The upper bound (equivalently lower bound on eigenvalues of Σ_0) guarantees that the compatibility condition (see Bühlmann and van de Geer 2011) is satisfied for the matrix $\Sigma_{-j,-j}^0$, which is the true covariance matrix Σ_0 without the jth row and jth column. The sub-Gaussianity condition A3 is used to obtain concentration results which are crucial to our analysis. Condition A3 is also used to ensure that the compatibility condition is satisfied for $\hat{\Sigma}$ with high probability (see Bühlmann and van de Geer 2011). Conditions A1, A2, and A3 are the same conditions as used in van de Geer et al. (2013) to obtain rates of convergence for the nodewise regression estimator.

Define the parameter set

$$\mathcal{G}(s) := \left\{ \Theta \in \mathbb{R}^{p\times p} : \max_{1\le i\le p} \|\Theta_i\|_0 \le s,\ A1 \text{ is satisfied} \right\}.$$

The following lemma shows that the proposed estimator $\hat{T}$ can be decomposed into a pivot term and a term which is of small order $1/\sqrt{n}$ with high probability.

Lemma 1 *Suppose that $\hat{\Theta}$ is the nodewise Lasso estimator with regularization parameters $\lambda_j \ge c\sqrt{\frac{\log p}{n}}$, uniformly in j, for some sufficiently large constant $c > 0$. Suppose that A2 and A3 are satisfied. Then, for each $(i, j) \in \mathcal{V} \times \mathcal{V}$, it holds:*

$$\sqrt{n}(\hat{T}_{ij} - \Theta^0_{ij}) = -\sqrt{n}(\Theta^0_i)^T(\hat{\Sigma} - \Sigma_0)\Theta^0_j + \Delta_{ij}, \tag{9}$$

where there exists a constant $C > 0$, such that

$$\lim_{n\to\infty} \sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P}\left(\max_{i,j=1,\ldots,p} |\Delta_{ij}| \ge C\frac{s\log p}{\sqrt{n}} \right) = 0.$$

From Lemma 1, it follows that we need to assume stronger sparsity condition than A2 for the remainder term Δ_{ij} to be negligible after normalization by $\sqrt{n}$. This is accordance with other literature on the topic, see van de Geer et al. (2013), Ren et al. (2015). Hence, we introduce the following strengthened sparsity condition.

A2* $\frac{s\log p}{\sqrt{n}} = o(1)$.

The next result shows that the elements of $\hat{T}$ are indeed asymptotically normal. To this end, we further define the asymptotic variance

$$\sigma^2_{ij} := \mathrm{var}\left(\left(\Theta^0_i\right)^T X_1 X_1^T \Theta^0_j \right).$$

In some of the results to follow, we shall assume a universal lower bound on σ_{ij} as follows.

A4 There exists a universal constant $\omega > 0$, such that $\sigma_{ij} \ge \omega$.

Assumption A4 is satisfied, e.g., under Gaussian design and A1. Denote a parameter set

$$\tilde{\mathcal{G}}(s) := \left\{ \Theta \in \mathbb{R}^{p\times p} : \max_{1\le i\le p} \|\Theta_i\|_0 \le s,\ A1,\ A4 \text{ are satisfied} \right\}.$$

Theorem 1 (Asymptotic normality) *Suppose that $\hat{\Theta}$ is the nodewise Lasso estimator with regularization parameters $\lambda_j \asymp \sqrt{\frac{\log p}{n}}$ uniformly in j. Suppose that A2* and A3 are satisfied. Then, for every $(i, j) \in \mathcal{V} \times \mathcal{V}$ and $z \in \mathbb{R}$, it holds*

$$\lim_{n\to\infty} \sup_{\Theta_0 \in \tilde{\mathcal{G}}(s)} |\mathbb{P}_{\Theta_0}\left(\sqrt{n}\left(\hat{T}_{ij} - \Theta^0_{ij}\right)/\sigma_{ij} \le z \right) - \Phi(z)| = 0.$$

To construct confidence intervals, a consistent estimator of the asymptotic variance σ_{ij} is required. Consistent estimators of σ_{ij} are discussed in Sect. 3.1 (see Lemmas 2 and 3). Hence, Theorem 1 implies uniformly valid asymptotic confidence intervals $I_\alpha := [\hat{T}_{ij} \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}_{ij}/\sqrt{n}]$, i.e.,

$$\lim_{n\to\infty} \sup_{\Theta_0 \in \tilde{\mathcal{G}}(s)} \left| \mathbb{P}_{\Theta_0}\left(\Theta^0_{ij} \in I_\alpha\right) - (1-\alpha)\right| = 0.$$

The result also enables testing hypotheses about individual elements of the precision matrix. For testing multiple hypothesis simultaneously, we may use the standard procedures, such as Bonferroni–Holm procedure (see van de Geer et al. 2013).

3.1 Variance estimation

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_0, as shown in the following lemma.

Lemma 2 *Suppose that assumptions A2 and A3 are satisfied and assume that the rows of the design matrix* $\mathbf{X}$ *are independent* $\mathcal{N}(0, \Sigma_0)$*-distributed. Let* $\hat{\Theta}$ *be the nodewise Lasso estimator and let* $\lambda_j \geq c\tau\sqrt{\log p/n}$ *uniformly in* j *for some* $\tau, c > 0$. *Then, for* $\hat{\sigma}^2_{ij} := \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}^2_{ij}$, *we have*

$$\sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P}\left(\max_{i,j=1,\ldots,p} \left|\hat{\sigma}^2_{ij} - \sigma^2_{ij}\right| \geq C_\tau\sqrt{s\log p/n}\right) \leq c_1 p^{1-\tau c_2},$$

for some constants $C_\tau, c_1, c_2 > 0$.

Lemma 2 implies that under $s = o(\sqrt{n}/\log p)$, we have a rate $|\hat{\sigma}^2_{ij} - \sigma^2_{ij}| = o_{\mathbb{P}}(1/n^{1/4})$. If Gaussianity is not assumed, we may replace the estimator of the variance with the empirical version, and plug in $\hat{\Theta}$ in place of the unknown Θ_0. Thus, we take the following estimator of σ^2_{ij}, where $\hat{\Theta}$ is the nodewise regression estimator:

$$\hat{\sigma}^2_{ij} := \frac{1}{n}\sum_{k=1}^{n}\left(\hat{\Theta}_i^T X_k X_k^T \hat{\Theta}_j\right)^2 - \hat{\Theta}^2_{ij}. \tag{10}$$

The following lemma justifies this procedure under A1, A2*, and A3.

Lemma 3 *Suppose that the assumptions A2* and A3 are satisfied and for some* $\epsilon > 0$, *it holds that* $\lim_{n\to\infty} \log^4(p \vee n)/n^{1-\epsilon} = 0$. *Let* $\hat{\Theta}$ *be the nodewise Lasso estimator and let* $\lambda_j \geq c\tau\sqrt{\log p/n}$ *uniformly in* j *for some* $\tau, c > 0$. *Let* $\hat{\sigma}_{ij}$ *be the estimator defined in* (10). *Then, for all* $\eta > 0$

$$\lim_{n\to\infty} \sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P}\left(\max_{i,j=1,\ldots,p} \left|\hat{\sigma}^2_{ij} - \sigma^2_{ij}\right| \geq \eta\right) = 0.$$

3.2 Rates of convergence

The desparsified estimator achieves optimal rates of convergence in supremum norm. Observe first that for the nodewise regression estimator, it holds by (7), Lemma 1 and Lemma 10 in the Online Resource 1 that

$$\hat{\Theta}_{ij} - \Theta^0_{ij} = \hat{\Theta}_i^T \left(\hat{\Sigma} \hat{\Theta}_j - e_j \right) + \mathcal{O}_{\mathbb{P}} \left(\max \left\{ \frac{1}{\sqrt{n}}, s \frac{\log p}{n} \right\} \right).$$

By Hölder's inequality and the KKT conditions, it follows

$$\left| \hat{\Theta}_i^T \left(\hat{\Sigma} \hat{\Theta}_j - e_j \right) \right| \leq \lambda_j \| \hat{\Theta}_i \|_1 / \hat{\tau}_j^2.$$

Consequently, for the rates of convergence of the nodewise Lasso in supremum norm, we find

$$\| \hat{\Theta} - \Theta_0 \|_\infty = \mathcal{O}_{\mathbb{P}} \left(\max \left\{ \max_{i,j=1,\ldots,p} \lambda_j \| \hat{\Theta}_i \|_1 / \hat{\tau}_j^2, \frac{1}{\sqrt{n}}, s \frac{\log p}{n} \right\} \right).$$

Desparsifying the estimator $\hat{\Theta}$ as in (8) removes the term involving $\lambda_j \| \hat{\Theta}_i \|_1 / \hat{\tau}_j^2$ in the above rates.

Theorem 2 (Rates of convergence) *Assume that A2 and A3 are satisfied. Let $\tau > 0$ and let $\hat{T}$ be the desparsified nodewise Lasso estimator with regularization parameters $\lambda_j \geq c\tau \sqrt{\log p / n}$ for some sufficiently large constant $c > 0$, uniformly in j. Then, there exist constants $C_\tau, c_1, c_2 > 0$, such that*

$$\sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P} \left(|\hat{T}_{ij} - \Theta^0_{ij}| \geq C_\tau \max \left\{ \frac{1}{\sqrt{n}}, s \frac{\log p}{n} \right\} \right) \leq c_1 e^{-c_2 \tau}.$$

and

$$\sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P} \left(\| \hat{T} - \Theta_0 \|_\infty \geq C_\tau \max \left\{ \sqrt{\frac{\log p}{n}}, s \frac{\log p}{n} \right\} \right) \leq c_1 p^{1 - c_2 \tau}.$$

We compare the results of Theorem 2 with results on optimal rates of convergence derived for Gaussian graphical models in Ren et al. (2015). Suppose that the observations are Gaussian, i.e., $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma_0)$. For $s \leq C_0 n / \log p$ for some $C_0 > 0$ and $p \geq s^\nu$ for some $\nu > 2$, it holds (see Ren et al. 2015)

$$\inf_{\hat{\Theta}_{ij}} \sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P} \left(|\hat{\Theta}_{ij} - \Theta^0_{ij}| > \max \left\{ C_1 \frac{1}{\sqrt{n}}, C_2 s \frac{\log p}{n} \right\} \right) > c_1 > 0,$$

and

$$\inf_{\hat{\Theta}} \sup_{\Theta_0 \in \mathcal{G}(s)} \mathbb{P} \left(\| \hat{\Theta} - \Theta_0 \|_\infty > \max \left\{ C_1' \sqrt{\frac{\log p}{n}}, C_2' s \frac{\log p}{n} \right\} \right) > c_2 > 0,$$

where C_1, C_2, C_1', C_2' are positive constants depending on ν and C_0 only. As follows from Theorem 2, the desparsified nodewise Lasso attains the lower bound on rates and thus is in this sense optimal (considering the Gaussian setting).

3.3 Other desparsified estimators

The desparsification may work for other estimators of the precision matrix, provided that certain conditions are satisfied. This is formulated in Lemma 4. A particular example of interest is the square-root nodewise Lasso estimator that will be discussed below. This estimator has the advantage that it is self-scaling in the variance, similarly as the square-root Lasso (Belloni et al. 2011) on which it is based.

Lemma 4 *Assume that for some estimator $\hat{\Omega} = (\hat{\Omega}_1, \ldots, \hat{\Omega}_p)$, it holds:*

$$\max_{j=1,\ldots,p} \|\hat{\Omega}_j - \Theta_j^0\|_1 = \mathcal{O}_{\mathbb{P}}(s\sqrt{\log p/n}), \ \ \|\hat{\Sigma}\hat{\Omega} - I\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}). \quad (11)$$

Then, for $\hat{T} := \hat{\Omega} + \hat{\Omega}^T - \hat{\Omega}^T\hat{\Sigma}\hat{\Omega}$, it holds under A1, A2, and A3*

$$\|\hat{T} - \Theta_0\|_\infty = \mathcal{O}_{\mathbb{P}}(\max\{s\log p/n, \sqrt{\log p/n}\}).$$

Moreover, $\sqrt{n}(\hat{T}_{ij} - \Theta_{ij}^0)/\sigma_{ij} \rightsquigarrow \mathcal{N}(0, 1)$.

We briefly consider nodewise regression with the square-root Lasso as an example. The square-root Lasso estimators may be defined via

$$\hat{\gamma}_j := \arg\min_{\gamma\in\mathbb{R}^{p-1}} \|X_j - \mathbf{X}_{-j}\gamma\|_2/n + 2\lambda_0\|\gamma\|_1,$$

for $j = 1, \ldots, p$. Define $\hat{\tau}_j^2 := \|X_j - \mathbf{X}_{-j}\hat{\gamma}_j\|_2^2/n$ and $\tilde{\tau}_j^2 := \hat{\tau}_j^2 + \lambda_0\hat{\tau}_j\|\hat{\gamma}_j\|_1$. The nodewise square-root Lasso is then given by $\hat{\Theta}_{j,\text{sqrt}} := \hat{\Gamma}_j/\tilde{\tau}_j^2$, where

$$\hat{\Gamma}_j := (-\hat{\gamma}_{j,1}, \ldots, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, \ldots, -\hat{\gamma}_{j,p})^T.$$

Note that compared with the nodewise Lasso, the difference lies in estimation of the partial correlations, where we used the square-root Lasso that “removes the square” from the squared loss. The Karush–Kuhn–Tucker conditions similarly as in Lemma 11 in Online Resource 1 give

$$\hat{\Sigma}\hat{\Theta}_{j,\text{sqrt}} - e_j - \frac{\hat{\tau}_j}{\tilde{\tau}_j^2}\lambda_0\tilde{Z}_j = 0,$$

where $\tilde{Z}_j := (\tilde{\kappa}_{j,1}, \ldots, \tilde{\kappa}_{j,j-1}, 1, \tilde{\kappa}_{j,j+1}, \ldots, \tilde{\kappa}_{j,p})^T$ and $\tilde{\kappa}_{j,i}$ is the subdifferential of the function $\beta \mapsto \|\beta\|_1$ with respect to β_i, evaluated at $\hat{\gamma}_j$. The paper Belloni et al. (2011) further shows that the ℓ_1-rates for the square-root Lasso satisfy condition (11).

The desparsified estimator may then be defined in the same way as in (8). Then, the conditions of Lemma 4 are satisfied, and this implies that a desparsified nodewise square-root Lasso achieves the same rates as the desparsified nodewise Lasso and thus is also rate-optimal.

3.4 The thresholded estimator and variable selection

The desparsified estimator can be used for variable selection without imposing irrepresentable conditions. Under mild conditions, the procedure leads to exact recovery of the coefficients that are sufficiently larger in absolute value than the noise level. The following corollary is implied by Theorem 2 and Lemmas 2 and 3.

Corollary 1 *Let $\hat{\Theta}$ be obtained using the nodewise Lasso and $\hat{T}$ be defined as in (8) with tuning parameters $\lambda_j \geq c\tau\sqrt{\log p/n}$ uniformly in j, for some $c, \tau > 0$. Assume that conditions A1, A2*, A3, and A4 are satisfied. Let $\hat{\sigma}_{ij}, i, j = 1, \ldots, p$ be the estimator from Lemma 3 and assume that $\log^4(p \vee n)/n^{1-\epsilon} = o(1)$ for some $\epsilon > 0$. Then, there exists some constant $C_\tau > 0$, such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i,j=1,\ldots,p} |\hat{T}_{ij} - \Theta^0_{ij}|/\hat{\sigma}_{ij} \geq C_\tau\sqrt{\log p/n}\right) = 0.$$

If, in addition, the rows of $\mathbf{X}$ are $\mathcal{N}(0, \Sigma_0)$-dsitributed and $\hat{\sigma}_{ij}, i, j = 1, \ldots, p$ is instead the estimator from Lemma 2, then there exist constants c_1, c_2, C_τ, such that

$$\mathbb{P}\left(\max_{i,j=1,\ldots,p} |\hat{T}_{ij} - \Theta^0_{ij}|/\hat{\sigma}_{ij} \geq C_\tau\sqrt{\log p/n}\right) \leq c_1 p^{1-c_2\tau}.$$

Corollary 1 implies that we may define the resparsified estimator

$$\hat{T}^{\text{thresh}}_{ij} := \hat{T}_{ij}\mathbf{1}_{|\hat{T}_{ij}|>C_\tau\hat{\sigma}_{ij}\sqrt{\log p/n}},$$

where $\hat{\sigma}_{ij}$ is defined as in Corollary 1. Denote $\hat{S}^{\text{thresh}} := \{(i, j) \in \mathcal{V}\times\mathcal{V} : \hat{T}^{\text{thresh}}_{ij} \neq 0\}$. Denote $S^{\text{act}}_0 := \{(i, j) \in \mathcal{V}\times\mathcal{V} : |\Theta^0_{ij}| \geq 2C_\tau\sigma_{ij}\sqrt{\log p/n}\}$. Then, it follows directly from Corollary 1 that with high probability

$$S^{\text{act}}_0 \subset \hat{S}^{\text{thresh}} \subset S_0.$$

The inclusion $S^{\text{act}}_0 \subset \hat{S}^{\text{thresh}}$ represents that $\hat{T}^{\text{thresh}}$ correctly identifies all the non-zero parameters which are above the noise level. The inclusion $\hat{S}^{\text{thresh}} \subset S_0$ means that there are no false positives. If for all $(i, j) \in S_0$ it holds:

$$|\Theta^0_{ij}| \geq 2C_\tau\sigma_{ij}\sqrt{\log p/n}, \tag{12}$$

then we have exact recovery, i.e., with high probability: $\hat{S}^{\text{thresh}} = S_0$.

4 Further comparison to previous work

Closely related is the paper Janková and van de Geer (2015), where asymptotically normal estimation of elements of the concentration matrix is considered based on the graphical Lasso. While the analysis follows the same principles, the estimation method used here does not require the irrepresentability condition (Ravikumar et al. 2008) that is assumed in Janková and van de Geer (2015). Hence, we are able to show that our results hold uniformly over the considered class. Furthermore, regarding the computational cost, our method uses Lasso regressions, which can be implemented using fast algorithms as in Efron et al. (2004). In comparison, the graphical Lasso method presents a more challenging computational problem, for more details see Mazumder and Hastie (2012).

Another related work is the paper Ren et al. (2015). This paper suggests an estimator for the precision matrix which is shown to have one dimensional asymptotically normal components with asymptotic variance to $\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2$. The assumptions and results used in the paper are essentially identical with our assumptions and theoretical results in the present paper. However, there are some differences. The paper Ren et al. (2015) assumes Gaussianity of the underlying distribution, while we only require sub-Gaussianity of the margins. Another difference is in the construction of the estimators. Both approaches use regression to estimate the elements of the precision matrix, but the paper Ren et al. (2015) concentrates on estimation of the joint distribution of each pair of variables (X^i, X^j) for $i, j = 1, \ldots, p$. Thus, it is computationally more intensive as it requires $\mathcal{O}(ps)$ high-dimensional regressions (see Ren et al. 2015), while our methodology only requires $\mathcal{O}(p)$.

5 Simulation results

In this section, we report on the performance of our method on simulated data and provide a comparison to another methodology. The random sample $X_1, \ldots, X_n$ satisfies $\mathbb{E}X_i = 0$, $\text{var}(X_i) = \Theta_0^{-1}$, where the precision matrix $\Theta_0 = \text{five-diag}(\rho_0, \rho_1, \rho_2)$ is defined by

$$\Theta^0_{ij} = \begin{cases} \rho_0 & \text{if } i = j, \\ \rho_1 & \text{if } |i - j| = 1, \\ \rho_2 & \text{if } |i - j| = 2, \\ 0 & \text{otherwise.} \end{cases}$$

We consider the settings $\mathbf{S}_1 = (\rho_0, \rho_1, \rho_2) = (1, 0.3, 0)$ and $\mathbf{S}_2 = (\rho_0, \rho_1, \rho_2) = (1, 0.5, 0.3)$. The second setting $(1, 0.5, 0.3)$ is further adjusted by randomly perturbing each non-zero off-diagonal element of Θ_0 by adding a realization from the uniform distribution on the interval $[-0.05, 0.05]$. We denote this new perturbed model by $(1, 0.5, 0.3)_U$. Hence, the second precision matrix was chosen randomly. The sparsity assumption requires $s = o(\sqrt{n}/\log p)$. We have chosen the sample sizes for numerical experiments according to the sparsity assumption (for this purpose, we ignored possible constants in the sparsity restriction), i.e., $n \geq s^2 \log^2 p$.

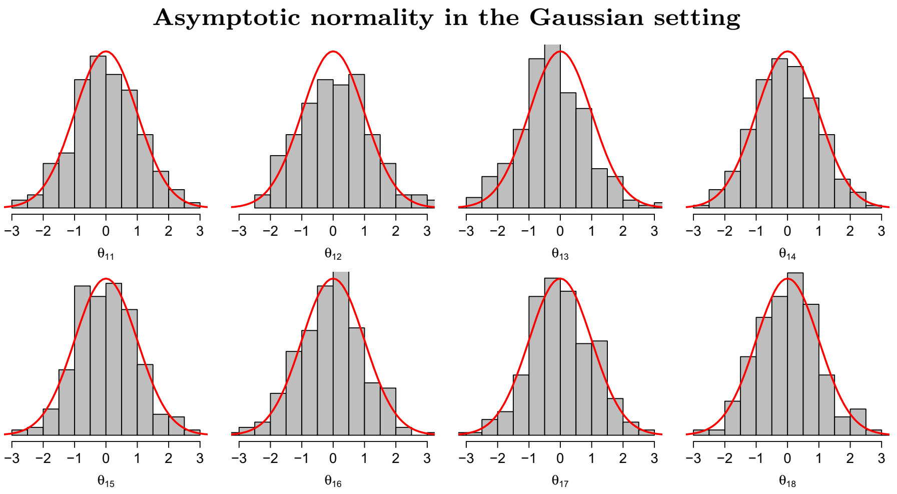

Fig. 1 Histograms of $\sqrt{n}(\hat{T}_{ij} - \Theta^0_{ij})/\hat{\sigma}_{ij}$, for $i = 1$, $j = 1, \ldots, 8$. The sample size was $n = 500$ and the number of parameters $p = 100$. The nodewise regression estimator was calculated 300 times. The setting is $\mathbf{S}_1 = (1, 0.3, 0)$

5.1 Asymptotic normality and confidence intervals for individual parameters

5.1.1 The Gaussian setting

In this section, we consider normally distributed observations, $X^i \sim \mathcal{N}(0, \Theta_0^{-1})$, for $i = 1, \ldots, n$. In Fig. 1, we display histograms of $\sqrt{n}(\hat{T}_{ij} - \Theta^0_{ij})/\hat{\sigma}_{ij}$ for $(i, j) \in \{(1, 1), (1, 2), (1, 3)\}$, where $\hat{T}$ is defined in (8) and the empirical variance $\hat{\sigma}_{ij}$ is estimated as suggested by Lemma 2. Superimposed is the density of $\mathcal{N}(0, 1)$.

Second, we investigate the properties of confidence intervals constructed using the desparsified nodewise Lasso. For comparison, we also provide results using confidence intervals based on the desparsified graphical Lasso introduced in Janková and van de Geer (2015). The coverage and length of the confidence interval were estimated by their empirical versions,

$$\hat{\alpha}_{ij} := \mathbb{P}_N \mathbf{1}_{\{\Theta_{0,ij} \in I_{ij,\alpha}\}} \text{ and } \hat{\ell}_{ij} := \mathbb{P}_N 2\Phi^{-1}(1 - \alpha/2)\hat{\sigma}_{ij}/\sqrt{n},$$

respectively, using $N = 300$ random samples. For a set $A \subset V \times V$, we define the average coverage over the set A (and analogously average length avglength$_A$) as

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

We report average coverages over the sets S_0 and S_0^c. These are denoted by avgcov$_{S_0}$ and avgcov$_{S_0^c}$, respectively. Similarly, we calculate average lengths of confidence intervals for each parameter Θ^0_{ij} from $N = 300$ iterations and report avglength_{S_0} and $\text{avglength}_{S_0^c}$.

The results of the simulations are shown in Tables 1 and 2. The target coverage level is 95 %. The methodology for the choice of the tuning parameters was used as follows (see Ren et al. 2015), for both methods:

$$\hat{s} = \sqrt{n}/\log p,\; B = \text{qt}(1 - \hat{s}/(2p), n-1),\; \lambda = B/\sqrt{n-1+B^2}, \qquad (13)$$

where $\text{qt}(\beta, n-1)$ denotes the β-quantile of a t-distribution with $n-1$ degrees of freedom.

5.1.2 A sub-Gaussian setting

In this section, we consider a design matrix with rows having a sub-Gaussian distribution other than the Gaussian distribution. Let $U := (U_1, \ldots, U_n)$ be an $n \times p$ matrix with jointly independent entries generated from a continuous uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$. Further consider a matrix $\Theta_0 := \text{five-diag}(1, 0.3, 0)$ and let $\Sigma_0 = \Theta_0^{-1}$. Then, we define

$$X_i := \Sigma_0^{1/2} U_i$$

for $i = 1, \ldots, n$. Then, the expectation of X_i is zero and the covariance matrix of X_i is exactly Σ_0 and the precision matrix is Θ_0. It follows by Hoeffding's inequality that X_i defined as above is sub-Gaussian with a universal constant $K > 0$.

A further difference compared to the simulations in Sect. 5.1.1 is that we now estimate the variance of the desparsified estimator using the formula proposed in (10) for sub-Gaussian settings:

Table 1 A table showing a comparison of desparsified nodewise Lasso (D–S NW) and desparsified graphical Lasso (D–S GL)

Setting $\mathbf{S}_1 = (1, 0.3, 0)$			S_0 avgcov	S_0 avglength	S_0^c avgcov	S_0^c avglength
p	n					
Gaussian setting: estimated coverage probabilities and lengths						
100	191	D–S NW	0.945	0.302	0.963	0.262
		D-S GL	0.931	0.293	0.974	0.254
200	253	D–S NW	0.947	0.267	0.963	0.232
		D-S GL	0.928	0.254	0.976	0.220
300	293	D–S NW	0.949	0.238	0.965	0.220
		D-S GL	0.928	0.236	0.977	0.205
400	324	D–S NW	0.948	0.246	0.965	0.230
		D-S GL	0.925	0.228	0.981	0.223

Parameter p takes values 100, 200, 300, 400 and the corresponding values n are given by $n = \lceil s^2 \log^2 p \rceil$, where $s = 3$. The regularization parameter was chosen as described in (13). The number of generated random samples was $N = 300$

Table 2 A table showing a comparison of desparsified nodewise Lasso (D-S NW) and the desparsified graphical Lasso (D-S GL)

Setting $\mathbf{S}_2 = (1, 0.5, 0.3)_U$			S_0 avgcov	S_0 avglength	S_0^c avgcov	S_0^c avglength
p	n					
Gaussian setting: estimated coverage probabilities and lengths						
100	531	D–S NW	0.896	0.164	0.975	0.146
		D-S GL	0.781	0.153	0.980	0.137
200	702	D–S NW	0.868	0.142	0.976	0.126
		D-S GL	0.729	0.133	0.982	0.119
300	814	D–S NW	0.863	0.131	0.976	0.117
		D-S GL	0.712	0.124	0.984	0.110
400	898	D–S NW	0.859	0.125	0.976	0.111
		D-S GL	0.709	0.118	0.984	0.105

Parameter p takes values 100, 200, 300, 400 and the corresponding values n are given by $n = \lceil s^2 \log^2 p \rceil$, where $s = 5$. The regularization parameter was chosen as described in (13). The number of generated random samples was $N = 300$

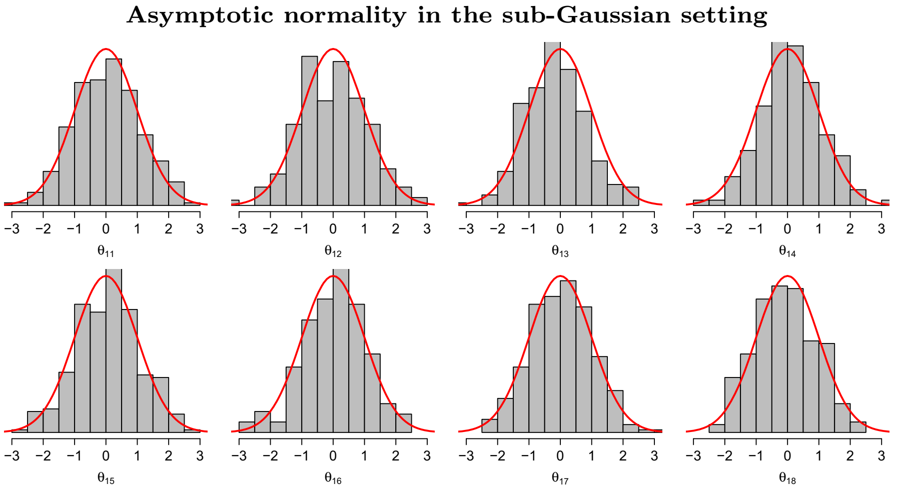

Fig. 2 Histograms of $\sqrt{n}(\hat{T}_{ij} - \Theta^0_{ij})/\hat{\sigma}_{ij}$, for $i = 1$, $j = 1, \ldots, 8$. The sample size was $n = 500$ and the number of parameters $p = 100$. The nodewise regression estimator was calculated 300 times. The setting is $\mathbf{S}_1 = (1, 0.3, 0)$

$$\hat{\sigma}^2_{ij} := \frac{1}{n} \sum_{k=1}^{n} (\hat{\Theta}_i^T X_k X_k^T \hat{\Theta}_j)^2 - \hat{\Theta}^2_{ij}, \tag{14}$$

where $\hat{\Theta}$ is the nodewise Lasso. The regularization parameters for the nodewise Lasso are used in accordance with (13). Figure 2 again shows the histograms related to several entries of the desparsified nodewise Lasso. Results related to the constructed

Table 3 A table showing a comparison of desparsified nodewise Lasso (D-S NW) and the desparsified graphical Lasso (D-S GL)

Setting $\mathbf{S}_1 = (1, 0.3, 0)$			S_0 avgcov	S_0 avglength	S_0^c avgcov	S_0^c avglength
p	n					
Sub-Gaussian setting: estimated coverage probabilities and lengths						
100	191	D–S NW	0.906	0.234	0.949	0.249
		D-S GL	0.811	0.190	0.944	0.216
200	253	D–S NW	0.909	0.203	0.950	0.217
		D-S GL	0.791	0.165	0.946	0.187
300	293	D–S NW	0.911	0.189	0.950	0.202
		D-S GL	0.765	0.152	0.947	0.173
400	324	D–S NW	0.911	0.180	0.951	0.192
		D-S GL	0.740	0.143	0.947	0.164

Parameter p takes values 100, 200, 300, 400 and the corresponding values n are given by $n = \lceil s^2 \log^2 p \rceil$, where $s = 3$. The regularization parameter was chosen as described in (13). The number of generated random samples was $N = 300$

confidence intervals are summarized in Table 3. The results demonstrate that the desparsified nodewise Lasso performs relatively well even under this non-Gaussian setting.

5.2 Variable selection

For variable selection as suggested in Corollary 1, we compare the desparsified nodewise Lasso and the desparsified graphical Lasso. The setting is again as in Section 5.1.1. Average true positives and false positives over 100 repetitions are reported. Choice of the tuning parameters is according to (13) and the thresholding level is given by

$$\lambda_{\text{thresh}} = \hat{\sigma}_{ij} \sqrt{2\nu \frac{\log p}{n}}, \tag{15}$$

taking $\nu = 1$ for the desparsified nodewise regression, $\nu = 0.5$ for the desparsified graphical Lasso. We take $\hat{\sigma}_{ij} = \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}_{ij}^2$ as in Lemma 2. The results of this simulation experiment are summarized in Table 4.

6 Real data experiments

We consider two real data sets, where we model the conditional independence structure of the covariates using a graphical model. In particular, we aim to do edge selection and we estimate the edge structure of the graphical model using the desparsified nodewise Lasso. The first data set is the Prostate Tumor Gene Expression data set, which is available in the `R` package `spls`. The second data set is about riboflavin (vitamin B_2) production by bacillus subtilis. The data set is available from the `R` package `hdi`.

Table 4 Estimated true positives (TP) and false positives (FP) for the desparsified nodewise regression estimator (D–S NW) and for the D–S GL estimator

Setting $\mathbf{S}_1 = (1, 0.5, 0.4)$		TP	TP rate %	FP	FP rate %
Estimated true positives (TP) and false positives (FP)					
$p = 100$	D–S NW	494	100.0	0	0
$\|S_0\| = 494$	D–S GL	493.98	99.999	0	0
$p = 200$	D-S NW	994	100.0	0	0
$\|S_0\| = 994$	D–S GL	993.62	99.961	0	0
$p = 300$	D–S NW	1494	100.0	0	0
$\|S_0\| = 1494$	D–S GL	1492.42	99.894	0	0
$p = 400$	D-S NW	1994.00	100.0	0	0
$\|S_0\| = 1994$	D–S GL	1989.08	99.753	0	0

The sample size $n = 400$ was held constant for all the values of p; the number of repetitions was $N = 100$. The thresholding levels was chosen as in (15)

For both data sets, the procedure is essentially identical. 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 observations. We calculate the nodewise Lasso using the tuning parameter as in the simulation study, and then calculate the desparsified nodewise Lasso. We threshold the desparsified nodewise Lasso at the level $\Phi^{-1}(1 - \alpha/(2p^2))\hat{\sigma}_{ij}/\sqrt{n}$, where $\alpha = 0.05$ and $\hat{\sigma}^2_{ij} = \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}^2_{ij}$ is an estimate of the asymptotic variance calculated under the assumption of normality and using the nodewise Lasso estimator $\hat{\Theta}$.

The first data set contained observations of $p = 4088$ logarithms of genes expression levels from $n = 71$ genetically engineered mutants of bacillus subtilis. We considered 500 variables with the highest variances; hence, a full graph contains $\binom{500}{2}$ edges. The desparsified nodewise Lasso identified 20 edges as significant. For comparison, the desparsified graphical Lasso introduced in Janková and van de Geer (2015) identified 5 edges as significant. It is worth pointing out that the set of edges selected by the desparsified graphical Lasso is a subset of the edges selected by desparsified nodewise Lasso.

The second data set contained $n = 102$ observations on $p = 6033$ variables. We used the procedure above to do edge selection using the desparsified nodewise Lasso. Our analysis identified 108 edges as significant using the desparsified nodewise Lasso. For comparison, the desparsified graphical Lasso identified 28 edges as significant. Again, the set of edges selected by the desparsified graphical Lasso is a subset of the edges selected by desparsified nodewise Lasso.

7 Conclusions

We proposed a methodology for low-dimensional inference in high-dimensional graphical models. The method, called the desparsified nodewise Lasso, is easy to implement and computationally competitive with the state-of-art methods. We studied asymptotic properties of the desparsified nodewise Lasso under mild conditions on the model. The desparsified nodewise Lasso enjoys rate optimality in supremum norm and leads to exact variable selection under beta-min conditions and mild conditions on the model. We demonstrated its performance on several models in a simulation study and on two real data sets. These numerical studies showed that it performs well in a variety of settings, including non-Gaussian settings. Further open questions concern, for instance, the asymptotic efficiency of the proposed estimator, similarly as in the low-dimensional settings.

References

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection amongst high-dimensional controls. Rev Econ Stud 81(2):608–650

Belloni A, Chernozhukov V, Wang L (2011) Square-root Lasso: Pivotal recovery of sparse signals via conic programming. Biometrika 98(4):791–806

Bickel PJ, Klaassen CA, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Springer, New York

Bickel PJ, Levina E (2008) Covariance regularization by thresholding. Ann Statist 36(6):2577–2604

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data. Springer, New York

Cai T, Liu W, Luo X (2011) A constrained l1 minimization approach to sparse precision matrix estimation. J Am Statist Assoc 106:594–607

Candes E, Tao T (2007) The dantzig selector: statistical estimation when p is much larger than n. Ann Statist 35(6):2313–2351

Chatterjee A, Lahiri SN (2011) Bootstrapping lasso estimators. J Am Statist Assoc 106(494):608–625

Chatterjee A, Lahiri SN (2013) Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. Ann Statist 41(3)

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Statist 32(2):407–451

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

Janková J, van de Geer S (2015) Confidence intervals for high-dimensional inverse covariance estimation. Electron J Statist 9:1205–1229

Javanmard A, Montanari A (2013) Model selection for high-dimensional regression under the generalized irrepresentability condition. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in neural information processing systems 26:3012–3020

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15(1):2869–2909

Knight K, Fu W (2000) Asymptotics for lasso-type estimators. Ann Statist 28(5):1356–1378

Lauritzen SL (1996) Graphical models. Clarendon Press, Oxford

Li KC (1989) Honest confidence regions for nonparametric regression. Ann Statist 17(3):1001–1008

Mazumder R, Hastie T (2012) The Graphical Lasso: New Insights and Alternatives. Electron J Statist, pp 2125–2149

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Statist 34(3):1436–1462

Ng B, Varoquaux G, P J-B, Thirion B (2013) A novel sparse group gaussian graphical model for functional connectivity estimation. Information Processing in Medical Imaging

Ravikumar P, Raskutti G, Wainwright MJ, Yu B (2008) High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. Electron J Statist 5:935–980

Ren Z, Sun T, Zhang C-H, Zhou HH (2015) Asymptotic normality and optimalities in estimation of large gaussian graphical models. Ann Statist 43(3):991–1026

Rothman AJ, Bickel PJ, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. Electron J Statist 2:494–515

Sun T, Zhang C-H (2012) Sparse matrix inversion with scaled Lasso. J Mach Learn Res 14:3385–3418

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58:267–288

van de Geer S (2016) Worst possible sub-directions in high-dimensional models. J Multi Anal 146:248–260

van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2013) On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Statist 42(3):1166–1202

van der Vaart A (2000) Asymptotic statistics. Cambridge University Press, Cambridge

Yuan M (2010) High dimensional inverse covariance matrix estimation via linear programming. J Mach Learn Res 11:2261–2286

Yuan M, Lin Y (2007) Model selection and estimation in the gaussian graphical model. Biometrika, page 117

Zhang C-H, Zhang SS (2014) Confidence intervals for low-dimensional parameters in high-dimensional linear models. J R Stat Soc Ser B Stat Methodol 76:217–242