Statistics Surveys
ISSN: 1935-7516

Testing for graph differences using the desparsified lasso in high-dimensional data

Lourens Waldorp

University of Amsterdam, Weesperplein 4, 1018 XA, the Netherlands
e-mail: waldorp@uva.nl

Abstract: Testing whether graphs are different is an essential tool in graph analysis for empirical sciences. For example, in neuroscience a graph could be obtained from a group with and without Alzheimers's disease. If the data are high-dimensional, that is, there are more parameters than observations, then a test statistic to determine whether graphs are different is not obvious. Natural extensions of Wald or score (type) tests, or likelihood ratio tests are problematic because (a) they require the covariance matrix of the parameters, which is inherently singular. Solutions exist, but are cumbersome. And (b), the Wald, score, and likelihood ratio tests' limiting distribution is chi-square, and so requires the degrees of freedom to be determined. This is far from trivial in high-dimensional settings. We propose, following Schott (2007), a test on the Frobenius norm of the difference between nodewise regression coefficients, which are used to obtain the graph parameters. We show that the statistic has an asymptotic standard normal distribution, and so has advantages that (a) no high-dimensional parameter covariance matrix needs to be inverted, and (b) no degrees of freedom need to be determined. We illustrate finite performance with some simulations.

Keywords and phrases: debiased lasso, large-scale graphs, high-dimensional inference.

1. Introduction

Comparison of graph structures (topologies) is becoming increasingly popular. For instance, in neuroscience typically a network is obtained in a cross-sectional design where different groups (or cohorts) of stages of Alzheimer are measured, so that changes in topology may lead to knowledge of the development of Alzheimer (Supekar, Musen and Menon, 2009). As another example, consider the comparison of a group of depressed and non-depressed subjects. A network can be constructed from the symptoms of the subjects in both groups (van Borkulo et al., 2014), which should then be compared in order to determine if the depressed and non-depressed differ with respect to the relations between symptoms.

In many situations the number of parameters exceeds the number of observations ($p > n$) such that standard likelihood or least squares estimation is not possible. Many algorithms have been put forward to deal with this high-dimensional situation (e.g., Meinshausen and Bühlmann, 2006; Friedman, Hastie

and Tibshirani, 2008; Sun and Zhang, 2012). Recently, interest is growing in statistical inference in high-dimensional statistics, by considering hypothesis testing (e.g., Meinshausen, Meier and Bühlmann, 2009; Bühlmann et al., 2013; Lockhart et al., 2014) and confidence intervals (e.g. Pötscher and Leeb, 2009; van de Geer et al., 2014; Javanmard and Montanari, 2014; Nickl et al., 2013). Here we use the desparsified lasso by van de Geer et al. (2014) and Javanmard and Montanari (2014) in a nodewise regression fashion to obtain the weights of a graph, for which the weights are meaningful. In van de Geer et al. (2014) and Javanmard and Montanari (2014) the asymptotic sample properties of the desparsified lasso are obtained by adding a component to the regular lasso and thereby allowing a standard analysis of asymptotic bias and variance of the estimator. Those results are used here to construct a hypothesis test to determine whether the weights in the graphs of the different groups are similar or not.

One way to determine differences between graphs is to construct a Wald type test, where the covariance matrix of the parameters is used to obtain a chi-square statistic under the null hypothesis (e.g., Muirhead, 1982; Bilodeau and Brenner, 1999). Of course, when $p > n$ the covariance matrix is singular and the Wald test needs correction (e.g., Andrews, 1987; Dufour and Valéry, 2011). A likelihood ratio test to determine graph differences cannot be used when $p > n$ (Schott, 2007), although Bai et al. (2009) obtained a valid correction using random matrix theory. This correction, however, is valid for testing sample covariance matrices, but does not seem a viable option for the current setting. Several alternatives have been proposed to test for the equality of sample covariance matrices when $p > n$ and when the data are (approximately) normal (Ledoit and Wolf, 2002; Srivastava, 2005; Schott, 2007). Here we use the idea of Schott (2007) to construct a test based on the Frobenius norm directly. We derive the mean and variance and then show normality by invoking the Häjek-Sidak central limit theorem. Two advantages of this test are that (a) there is no need to construct a generalised or otherwise regularised inverse of the parameter covariance matrix, and (b) there is no need to specify the degrees of freedom, which is often difficult when $p > n$.

2. Undirected graphical models

An undirected graphical model or Markov random field is a set of probability distributions representing the structure of a graph G. Let $G = (V, E)$ be an undirected graph, where V is the set of nodes $\{1, 2, \ldots, p\}$ and $E = V \times V$ is the set of edges $\{(s, t) : s, t \in V\}$, with size $|E| = m$. We associate with each vertex $s \in V$ a random variable X_s. For any subset $A \subset V$ of nodes in we define a configuration $x_A = \{x_s : s \in A\}$; we often write $x_{\setminus s}$ to mean $x_{V \setminus \{s\}}$. For subsets of nodes A, B, and W, we denote by $X_A \perp\!\!\!\perp X_B \mid X_W$ that X_A is conditionally independent of X_B given X_W. A random vector X is Markov with respect to G if $X_A \perp\!\!\!\perp X_B \mid X_W$ whenever removing W creates two disjoint subsets A and B. A clique C is a (sub)set of nodes such that any pair of nodes in C has an edge. For each clique C in the set of all cliques $\mathcal{C}$ of graph G a potential

function $\psi_C : \mathcal{X}^{|C|} \to \mathbb{R}_+$ maps the states of the nodes in clique C to the positive reals. When normalized, the product of the potential functions defines the distribution. The distribution of the random vector Z factorizes according to graph G if it can be represented by a product of potential functions of the cliques

$$p(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{1}$$

For strictly positive distributions the Hammersly-Clifford theorem says that the Markov and factorization properties are equivalent (see, e.g., Cowell et al., 1999; Lauritzen, 1996).

Consider the example of a Gaussian random field. Let $X \in \mathcal{X}^p = \mathbb{R}^p$ be a continuous random vector associated with the graph $G = (V, E)$. If we assume a multivariate Gaussian (normal) distribution for X with mean μ and covariance Σ, then the usual form of the distribution is

$$p_\theta(x) = c_\theta \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$$

It can be shown that the Markov property results in zeros of the precision (inverse covariance) matrix (Lauritzen, 1996). If $\Theta_{st} = 0$ in the precision matrix $\Theta = \Sigma^{-1}$, then s and t are conditionally independent given all other variables and $(s,t) \notin E$. In other words, $\Theta_{st} = 0$ whenever $X_s \perp\!\!\!\perp X_t \mid X_{V\setminus\{t,s\}}$.

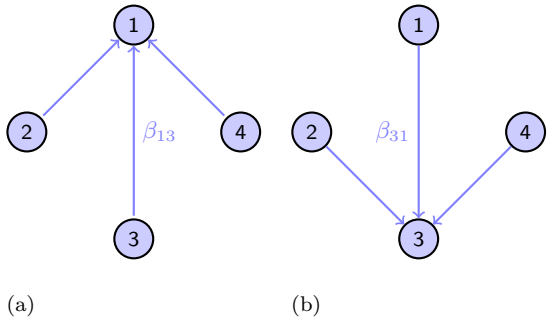
3. Estimation of graph parameters

Meinshausen and Bühlmann (2006) use a result of Lauritzen (1996) to show that identifying the edges in a Gaussian Markov random field is identical to determining the neighborhood of each of the p nodes. Let X be a Gaussian random field of dimension p with mean 0 and covariance Σ. As seen in the example in the previous section, the inverse $\Theta = \Sigma^{-1}$ is by the Hammersly-Clifford theorem indicative of conditional independence. Hence, the neighborhood of any node $s \in V$ in the Gaussian random field can be determined from the zeros in the vector $(\Theta_{st}, t \in V\setminus s)$. In fact, the Markov property gives for any nodes in A and B such that $\Theta_{AB} = 0$

$$X_A \perp\!\!\!\perp X_B \mid X_{V\setminus A,B}$$

It follows that we can define a neighorhood of any node s in terms of the nonzero values in Θ, i.e., $\text{ne}(s) = \{t \in V\setminus s : \Theta_{st} \neq 0\}$. The regression coefficients in the $p-1$ vector β_s obtained from regressing X_s on the remaining nodes $X_{V\setminus s}$, and is $\beta_{st} = -\Theta_{st}/\Theta_{ss}$ (Lauritzen, 1996). Hence the neighborhood of s can be determined in terms of regression coefficients

$$\text{ne}(s) = \{t \in V\setminus s : \beta_{st} \neq 0\} \tag{2}$$



(a) (b)

FIG 1. *Graphs of regressions involving the nodes* 1 *and* 3 *twice.*

Since each combination of nodes occurs twice in such a series of regressions (see Figure 1), β_{st} and β_{ts} have to be combined by an *or* rule, where either can be nonzero, or an *and* rule, where both have to be nonzero. In other words, a series of regressions leads to an estimate of the inverse covariance matrix $\hat{\Theta}$.

The parameters of a Markov random field can be obtained in several ways (see, e.g., Bühlmann and van de Geer, 2011). One of the most popular choices, which we will focus on here, is the nodewise regression approach, introduced by Meinshausen and Bühlmann (2006). Let $X \in \mathbb{R}_n^p$ be a Gaussian random variable with mean 0 and covariance Σ. For each $s \in V$, assign $Y_s = X_s$, the sth column of X, and obtain a lasso estimate $\hat{\beta}_{L,s}$ for the sth node in X by minimizing

$$\tau^2_{\lambda_s}(\beta_s) = (Y_s - X_{\setminus s}\beta_s)^\mathsf{T}(Y_s - X_{\setminus s}\beta_s)/n + \lambda_s||\beta_s||_1 \tag{3}$$

where $||\beta_s|| = \sum_i |\beta_{s,i}|$ is the ℓ_1 norm. Each vector $\hat{\beta}_{L,s} = (\hat{\beta}_{L,s,i}, i \in V\setminus s)$ has $p-1$ elements, giving the regression coefficients of all remaining nodes on s. Since each coefficient is $\beta_{st} = -\Theta_{st}/\Theta_{ss}$, we can obtain the inverse covariance matrix by multiplying by Θ_{ss}. Let $\hat{\tau}^2_s = \tau^2_{\lambda_s}(\hat{\beta}_{L,s})$ be the estimate of $1/\Theta_{ss}$. The completed $p \times p$ matrix is then

$$\hat{\Theta}_{G,st} = \begin{cases} -\beta_{L,st}/\hat{\tau}^2_s & \text{if } s \neq t \\ 1/\hat{\tau}^2_s & \text{if } s = t \end{cases} \tag{4}$$

There are many algorithms to obtain $\hat{\beta}_{L,s}$. Here we focus on the desparsified lasso because it is appealing to use for inference.

3.1. Nodewise regression with the desparsified lasso

The starting point of the desparsified lasso is obtaining the 'normal equations' as if a regular least squares solution could be obtained. Then van de Geer et al. (2014) show that the remainder is negligible as sample size increases. We can obtain a 'normal equation' as with least squares, with the difference that we use

a subgradient for the ℓ_1 norm $||\beta_s||_1$. Since this is not differentiable at $\hat{\beta}_{s,i} = 0$, we have that $\partial|\hat{\beta}_{s,i}| = \text{sign}(\hat{\beta}_{s,i})$ when $\hat{\beta}_{s,i} \neq 0$ and $\partial|\hat{\beta}_{s,i}|$ is the interval $[-1, 1]$ when $\hat{\beta}_{s,i} = 0$. The subgradient is

$$-2X_{\setminus s}^{\mathsf{T}}(Y_s - X_{\setminus s}\hat{\beta}_{L,s})/n + 2\lambda_s \partial||\hat{\beta}_{L,s}||_1 = 0$$

where $\hat{\beta}_{L,s}$ is some lasso estimate (e.g., Meinshausen-Bühlmann version) and $\partial||\hat{\beta}_{L,s}||_1$ represents the subdifferential. This set of conditions for optimization is called the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004). Let $\hat{\Sigma}_{\setminus s,\setminus s} = X_{\setminus s}^{\mathsf{T}} X_{\setminus s}/n$, which is the covariance matrix without row and column s. Rewriting these conditions gives

$$\hat{\Sigma}_{\setminus s,\setminus s}(\hat{\beta}_{L,s} - \beta_s) + \lambda \partial||\hat{\beta}_{L,s}||_1 = X_{\setminus s}^{\mathsf{T}}\varepsilon_s/n$$

Now all we need to do is get rid of the term $\hat{\Sigma}_{\setminus s,\setminus s}$ and we obtain a way to get to β_s. Unfortunately, this is impossible because in general $\hat{\Sigma}_{\setminus s,\setminus s}$ is not of full rank, it has rank $\min(n, p)$, which could very well be n. We could use an approximate inverse $\hat{\Theta}_s$ of $\hat{\Sigma}_{\setminus s,\setminus s}$ so that

$$\hat{\beta}_{L,s} - \beta_s + \hat{\Theta}_s \lambda_s \partial||\hat{\beta}_{L,s}||_1 = \hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}\varepsilon_s/n - \Delta_s/\sqrt{n} \tag{5}$$

and

$$\Delta_s = \sqrt{n}(\hat{\Theta}_s \hat{\Sigma}_{\setminus s,\setminus s} - I_{p-1})(\hat{\beta}_{L,s} - \beta_s)$$

van de Geer et al. (2014, section 5) show that under several assumptions (see below) Δ_s is negligible. Then a reasonable way to get rid of the bias (or desparsifying the lasso) is to use

$$\hat{\beta}_{dL,s} = \hat{\beta}_{L,s} + \hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}(Y_s - X_{\setminus s}\hat{\beta}_{L,s})/n \tag{6}$$

In van de Geer et al. (2014) the choice for the approximate inverse is the Meinshausen-Bühlmann nodewise regression lasso, since in that case a clear bound can be obtained on Δ_s which makes it work. In the Appendix we show that also the shrinkage estimator of Ledoit and Wolf (2004) has such a bound. A shrinkage estimator using weight $0 \leq \rho \leq 1$ is $\hat{\Sigma}_{\setminus s,\setminus s,\rho} = \rho\mu I_{p-1} + (1 - \rho)\hat{\Sigma}_{\setminus s,\setminus s}$. Javanmard and Montanari (2014) show that an algorithm to obtain $\hat{\Theta}_s$ which minimises the variance of the desparsified lasso subject to bounding $||\hat{\Theta}_s \hat{\Sigma}_{\setminus s,\setminus s} - I_{p-1}||_\infty$ results in the desired properties of the desparsified lasso. In the main text we work with the approach of van de Geer et al. (2014) where the Meinshausen and Bühlmann (2006) nodewise regression is used to determine the approximate inverse. The assumptions underlying the nodewise regressions with the desparsified lasso for Gaussian data are the following.

Assumption 1 (True model) Let $G = (V, E)$ be an undirected graph represented in a multivariate normal distribution with mean 0 and covariance Σ. Let

the rows of X be independent draws from $N_p(0, \Sigma)$. The nodewise regression then assumes that for $Y_s = X_s$ and all $s \in V$,

$$Y_s = X_{\setminus s}\beta_s + \varepsilon_s$$

with random $n \times (p-1)$ design matrix $X_{\setminus s}$, an unknown $p-1$ vector of parameters β_s, and residuals ε_s that are normally distributed with mean 0 and covariance $\sigma^2 I_n$ which are independent of $X_{\setminus s}$ and where $\sigma^2 < \infty$.

The assumption is strong in the sense that the linear structure should hold exactly (but see Bühlmann and van de Geer (2011, chap. 6) for extensions to nonlinear variants in the lasso).

Assumption 2 (Properties of Σ) The matrix Σ has positive smallest eigenvalue $\delta_{\min} > 0$ such that $1/\delta_{\min} = O(1)$, and the largest variance $\max_{s \in V} \Sigma_{ss}$ is bounded.

This assumption is slightly stronger than in van de Geer et al. (2014) since we will use this assumption also for nodewise regression using the desparsified lasso. Assumption 2 implies that the same holds for each principal submatrix (deleting a row and corresponding column), and so implies the assumption in van de Geer et al. (2014).

Assumption 3 (Sparsity) It is assumed that the number of nonzero edges s_0 to each node is sparse, i.e., the sparsity of each node (or row of Σ^{-1}) s_0 is assumed to be of order $o(\sqrt[3]{n}/\log(p-1))$.

This sparsity assumption is slightly stronger that the one in van de Geer et al. (2014), which is $o(\sqrt{n}/\log(p-1))$. This is because of the additional term $\hat{\Theta}_s$ which is of order $O_p(\sqrt{s_0})$ in the ℓ_2 norm that we use here to determine a test statistic (see Corollary 2 in Section 6). With slightly more relaxed assumptions van de Geer et al. (2014) prove the following.

Theorem 1 Suppose the linear model from Assumption 1 holds, and the properties of Σ in Assumption 2 and the sparsity Assumption 3 are both satisfied. Consider that the approximate inverse $\hat{\Theta}_s$ is obtained by nodewise regression with penalty of order $O(\sqrt{\log(p-2)/n})$. Then for $\lambda_s = O(\sqrt{\log(p-1)/n})$ in (3) has negligible Δ_s and

$$\sqrt{n}(\hat{\beta}_{dL,s} - \beta_s)|X_{\setminus s} \sim N_{p-1}(0, \sigma^2 \hat{\Theta}_s \hat{\Sigma}_{\setminus s,\setminus s} \hat{\Theta}_s^{\mathsf{T}}) \tag{7}$$

and $||\hat{\Theta}_s \hat{\Sigma}_{\setminus s,\setminus s} \hat{\Theta}_s^{\mathsf{T}} - \Theta_s||_\infty = o_p(1)$.

For each node $s \in V$ we use the estimate $\hat{\beta}_{dL,s}$, scaled by $\tau_s^2(\hat{\beta}_L)$ and then obtain the estimate $\hat{\Theta}_G$ in (4). Theorem 1 then means we can apply standard

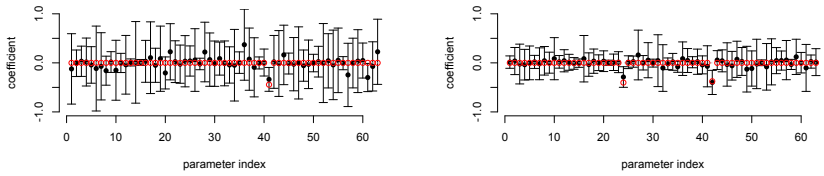


FIG 2. *Estimates (black, filled circles) using the desparsified lasso (6) and their confidence bounds (black lines) estimated by (7). The true values are the red, open circles. A single regression is shown with 63 predictors. In the left panel are the estimates and bounds obtained from a simulation with* $n = 50$*, and in the right panel are the estimates and bounds obtained from a simulation with* $n = 100$ *(see Section 5 for details).*

theory to obtain confidence intervals for each parameter i conditional on $X_{\setminus s}$

$$[\hat{\beta}_{dL,s,i} - c_\alpha, \hat{\beta}_{dL,s,i} + c_\alpha]$$

where

$$c_\alpha = \Phi^{-1}(1-\alpha/2)\sigma\sqrt{(\hat{\Theta}_{\setminus s}\hat{\Sigma}_{\setminus s,\setminus s}\hat{\Theta}_{\setminus s}^{\mathsf{T}})_{ii}/n}$$

and Φ^{-1} is the inverse standard normal cumulative function. This could be used to select edges by considering whether or not 0 is in the interval (variable selection).

Obviously, we would like to conclude that all regressions for all nodes $s \in V$ are correct, so that the true underlying graph is recovered. This follows almost immediately from Theorem 1.

Corollary 2 Let $\hat{\Theta}_G$ be obtained with the desparsified lasso with λ_s of order $O(\sqrt{\log(p-1)/n})$ for all $s \in V$. Suppose Assumptions 1-3 hold. Then

$$||\hat{\Theta}_G - \Sigma^{-1}||_\infty = o_p(1) \tag{8}$$

A proof is given in Section 6. As a small illustration, in Figure 2 the estimates and their confidence bounds are shown, obtained from a simulation (see Section 5 for details). The left panel is obtained with 50 observations and the right panel with 100. The estimates (black circles) are close to the true values (red, open circles), for both zero and nonzero coefficients, and the confidence intervals obtained with (7) often contain the true values.

Our aim is to use these results to be able to compare graphs (differences in edges) between groups in terms of hypothesis tests.

4. Comparing graphs

Close to the idea of tests on squared residuals, we use the idea popular in testing sample covariance matrices of independent groups (Jöreskog and Sörbom,

1989; Muirhead, 1982). We follow Schott (2007) and use the idea of obtaining the asymptotic distribution of a random variable with squared differences between populations (Frobenius norm). The advantage of this statistic is that the parameter covariance matrix is not used, as it is in the Wald (type) statistic, avoiding issues with nonsingularity (Andrews, 1987). Furthermore, the distribution under the null hypothesis is standard normal so that there are no degrees of freedom, which can be problematic in high-dimensional data (Bühlmann and van de Geer, 2011).

We want to test whether the population graphs are equivalent. We therefore have the hypotheses

$$H_0 : \Theta_1 = \Theta_2 \quad \text{vs} \quad H_a : \Theta_1 \neq \Theta_2$$

Instead of working directly with the partial covariance matrix $\hat{\Theta}_G$, we use the estimators from the regression $\hat{\beta}_{dL,s}$ of dimension $p-1$, which are the rows of the $p \times p$ matrix $\hat{\Theta}_G$ without the diagonal element. We can collect all rows in the $p \times (p-1)$ matrix $\hat{B}$.

We start with the simple idea of the test, which is the Frobenius norm of the difference of the nodewise regression coefficients

$$||\hat{B}_1 - \hat{B}_2||_F^2 = \sum_{s \in V} \sum_{t \in V} (\hat{\beta}_{1,dL,st} - \hat{\beta}_{2,dL,st})^2 \tag{9}$$

We derive the asymptotic mean and variance of our statistic and show that it converges to a standard normal random variable if properly scaled.

We rewrite the the difference between the estimates of the groups

$$\hat{\beta}_{1,dL,s} - \hat{\beta}_{2,dL,s} = (\beta_{1,s} - \beta_{2,s}) + [(\hat{\beta}_{1,dL,s} - \beta_{1,s}) - (\hat{\beta}_{2,dL,s} - \beta_{2,s})]$$

We can then use this to obtain the asymptotic expectation of the Frobenius norm

$$\begin{aligned} &\mathbb{E}[||\hat{B}_1 - \hat{B}_2||_F^2 \mid X_{j,\setminus s}, s \in V; j = 1,2] = \\ &\qquad \sum_{s \in V} ||\beta_{1,s} - \beta_{2,s}||_2^2 + \mathbb{E}||\hat{\beta}_{1,dL,s} - \beta_{1,s}||_2^2 + \mathbb{E}||\hat{\beta}_{2,dL,s} - \beta_{2,s}||_2^2 \end{aligned} \tag{10}$$

where we take conditional expectations separately for each node $s \in V$ given $X_{j,\setminus s}$. If we therefore define the random variable

$$\hat{m}_{1,2} = ||\hat{B}_1 - \hat{B}_2||_F^2 - \sum_{s \in V} \mathbb{E}||\hat{\beta}_{1,dL,s} - \beta_{1,s}||_2^2 - \mathbb{E}||\hat{\beta}_{2,dL,s} - \beta_{2,s}||_2^2$$

then

$$\mathbb{E}[\hat{m}_{1,2} \mid X_{j,\setminus s}, s \in V; j = 1,2] = \sum_{s \in V} ||\beta_{1,s} - \beta_{2,s}||_2^2 + o(1)$$

and so is 0 only if H_0 is true. The desparsified lasso allows computation of the expectation and variance by considering Corollary 2. There we find that

$$||\hat{\beta}_{dL,s} - \beta_s||_2^2 = ||\hat{\Theta}_s X_{\setminus s}^\mathsf{T} \varepsilon_s / n||_2^2 + o_p(1)$$

Then we have that for each component conditionally on $X_{\setminus s}$

$$\mathbb{E}[||\hat{\beta}_{j,dL,s} - \beta_{1,s}||_2^2] = \sigma_j^2 \text{tr}(\hat{\Omega}_{j,s})/n_j \tag{11}$$

$$\mathbb{V}[||\hat{\beta}_{j,dL,s} - \beta_{1,s}||_2^2] = \sigma_j^4 \text{tr}(\hat{\Omega}_{j,s})^2/n_j^2 \tag{12}$$

It follows that the variance of $\hat{m}_{12}$ is

$$\hat{\psi}_{12} = \sum_{s \in V} \sigma_1^4 \text{tr}(\hat{\Omega}_{1,s})^2/n_1^2 + \sigma_2^4 \text{tr}(\hat{\Omega}_{2,s})^2/n_2^2 + \sigma_1^2 \sigma_2^2 \text{tr}(\hat{\Omega}_{1,s}) \text{tr}(\hat{\Omega}_{2,s})/(n_1 n_2)$$

Then we can construct the statistic $T_{n,p}(X_1, X_2) = \hat{m}_{12}/\sqrt{\hat{\psi}_{12}}$ to test the difference in the coefficients of all nodewise regressions. The test statistic can be written as

$$T_{n,p}(X_1, X_2) = \frac{||\hat{B}_1 - \hat{B}_2||_F^2 - \hat{\sigma}_1^2 \sum_{s \in V} \text{tr}(\hat{\Omega}_{1,s})/n_1 - \hat{\sigma}_2^2 \sum_{s \in V} \text{tr}(\hat{\Omega}_{2,s})/n_2}{\sqrt{\sum_{s \in V} \hat{\sigma}_1^4 \text{tr}(\hat{\Omega}_{1,s})^2/n_1^2 + \hat{\sigma}_2^4 \text{tr}(\hat{\Omega}_{2,s})^2/n_2^2 + \hat{\sigma}_1^2 \hat{\sigma}_2^2 \text{tr}(\hat{\Omega}_{1,s}) \text{tr}(\hat{\Omega}_{2,s})/(n_1 n_2)}} \tag{13}$$

where we use the estimate

$$\hat{\sigma}_j^2 = \frac{1}{p} \sum_{t=1}^{p} \left(\frac{1}{n} \sum_{s=1}^{n} (Y_{j,st} - \bar{Y}_{j,t})^2 \right) \tag{14}$$

which is consistent under Assumption 1 (Ferguson, 1996). Here we use the Hájek-Sidak central limit theorem to prove that under H_0 (see Section 6)

$$T_{n,p}(X_1, X_2) \xrightarrow{d} N(0,1)$$

This idea can be generalised to q groups. For each of the $q(q-1)/2$ comparisons, if the null hypothesis $H_0 : B_1 = B_2 = \cdots = B_q$ is true, then we have $q(q-1)/2$ $N(0,1)$ random variables. Hence, we can scale a sum of all $T_{n,p}(X_i, X_j)$ by $1/\sqrt{q(q-1)/2}$ to again obtain a standard normal variable.

Theorem 3 Let the matrix with coefficients $\hat{B}_j$ for group j be the desparsified lasso estimates of (6). Suppose that Assumptions 1-3 hold and that the groups $j = 1, 2, \ldots, q$ are independent with $\text{tr}(\Sigma_j)^2 = O(1)$. Then under $H_0 : B_1 = \cdots = B_q$, the statistic $q(q-1)/2)^{-1/2} \sum_{i<j} T_{n,p}(X_i, X_j)$ converges in distribution to $N(0,1)$.

TABLE 1
Averaged values (across simulations) of the true difference between regression coefficients d_{12}, the estimated difference $\hat{d}_{12}$, and the standard error $\sqrt{\hat{\psi}_{12}}$ when H_a is true and there is a difference in graphs.

	n/p				
	0.781	1.563	3.125	4.688	6.25
d_{12}	29.631	29.641	29.437	30.636	28.861
$\hat{d}_{12}$	65.318	44.262	34.840	32.052	29.935
$\sqrt{\hat{\psi}_{12}}$	62.044	32.903	14.938	9.279	7.937

A proof is given in Section 6. It is obvious from the statistic (13) that the probability of rejecting H_0 depends on the size of the variance, that is,

$$\mathbb{P}[T_{n,p}(X_1, X_2) \geq c_\alpha] \to 1 - \mathbb{P}\left[\sqrt{\hat{\psi}_{12}} c_{\alpha/2} + \hat{\psi}_1 + \hat{\psi}_2 \geq \sum_{s \in V} ||\beta_{1,s} - \beta_{2,s}||_2^2\right]$$

where $c_{\alpha/2}$ is the upper $\alpha/2$-tail of the standard normal distribution, and $\hat{\psi}_j = \sigma_j^2 \sum_{s \in V} \text{tr}(\hat{\Omega}_{j,s})/n_j$.

As such a test is intended to be used in practice, it is of interest to know how well it performs in finite samples.

5. Simulations

We simulated random networks with $p = 64$ nodes and an expected number of edges of 50.4 (2.5%). Fixed weights were placed on the nonzero edges with values ranging from 0.2 to 0.8. From this graph we generated normally distributed data for $n = 50$, 100, 200, 300, and 500 observations. For the generation of data we used the `R` package `huge` (Zhao et al., 2012), and for the estimation of networks we used `glmnet` (Friedman, Hastie and Tibshirani, 2010). The tuning parameters were all obtained with 10-fold cross validation. We used 100 simulations to determine the relevant statistics.

We give results for the ratios n/p, which range approximately from 0.781 to 7.813. These settings ensured that the sparsity assumption on each node $o(\sqrt[3]{n}/\log(p-1))$ is almost never violated; the expected degree of a node is 0.788 and with $n = 50$, and nodewise sparsity is 0.889. Data were generated independently and identically distributed in samples of size n with mean 0 and inverse covariance matrix Θ_j for $j = 1, 2$ groups. The groups were either the same under H_0, where the Frobenius norm was 0, or different under H_a, where the average Frobenius norm is given in Table 1. Under H_0 the difference in norm for the *estimated* parameters was on average 19.823, and this difference under H_a was is also given in Table 1.

Variable selection was used to determine nonzero edges in a graph: Each edge was tested for significance at the Bonferroni level $\alpha/(p(p-1))$, with $\alpha = 0.05$. We used this correction since we tested all $p(p-1)$ coefficients separately. Based on

these thresholded coefficients, recall (true positive rate) and precision (positive predictive value) were determined. For the true edge set E and estimated edge set $\hat{E}$, recall and precision are defined as

$$\text{recall} = \frac{\hat{E} \cap E}{E} \qquad \text{precision} = \frac{\hat{E} \cap E}{\hat{E}}$$

The average coverage of the confidence intervals is determined across edges within (and averaged across simulations), separately for both nonzero coefficients s_0 and zero coefficients s_0^c. They are defined as

$$\text{coverage}(s_0) = \frac{1}{|E|} \sum_{(s,t) \in E} \mathbb{P}[\Theta_{st} \in \text{CI}(\hat{\Theta}_{st})]$$

$$\text{coverage}(s_0^c) = \frac{1}{|E^c|} \sum_{(s,t) \in E^c} \mathbb{P}[\Theta_{st} \in \text{CI}(\hat{\Theta}_{st})]$$

The average length of the confidence intervals is determined similarly

$$\text{length}(s_0) = \frac{1}{|E|} \sum_{(s,t) \in E} \text{length}[\text{CI}(\hat{\Theta}_{st})]$$

$$\text{length}(s_0^c) = \frac{1}{|E^c|} \sum_{(s,t) \in E^c} \text{length}[\text{CI}(\hat{\Theta}_{st})]$$

5.1. Results

Figure 3 shows that recall is reasonably high with ratio $n/p = 0.781$ (50 observations), and is nearly 1 when the ratio is 3.125 (200 observations). Additionally, the precision is reasonably high, so that nearly all edges that are found significant are indeed correctly recovered edges. In Figure 4 the coverage and length of the confidence intervals computed with (7) are shown. The coverage of the nonzero coefficients is around 0.8 for each ratio n/p, while the coverage is near 1 for the zero coefficients. The length of the interval decreases as expected with the increasing ratio n/p.

The performance of the test $T_{n,p}(X_1, X_2)$ is shown in Figure 5. The left panel shows that the false positive rate remains under or around the nominal 5% level for all levels of observations per parameter. The right panel shows that the power only starts to increase when there are at least 7 observations per parameter but is at 1 when the this ratio is 10.

6. Proofs

Proof (Corollary 2) Comparing the supremum and ℓ_2 norm gives

$$||\hat{\Theta}_G - \Theta||_\infty \leq \max_{s \in V} ||\hat{\Theta}_{G,s} - \Theta_s||_2$$

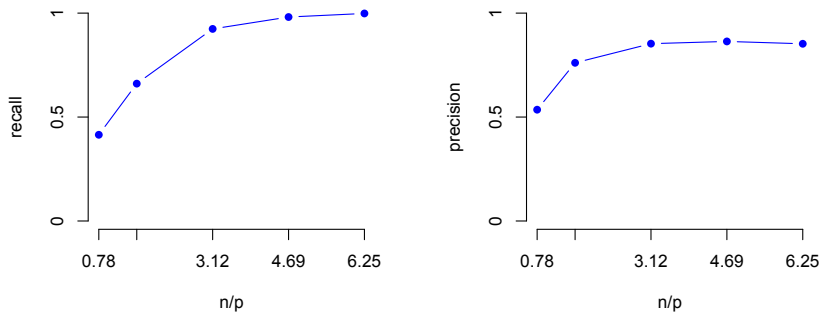


FIG 3. *Recall (left panel) and precision (right panel) for different numbers of observations per parameter n/p, with $p = 64$ and $n = 50$, 100, 200, 300, and 400.*

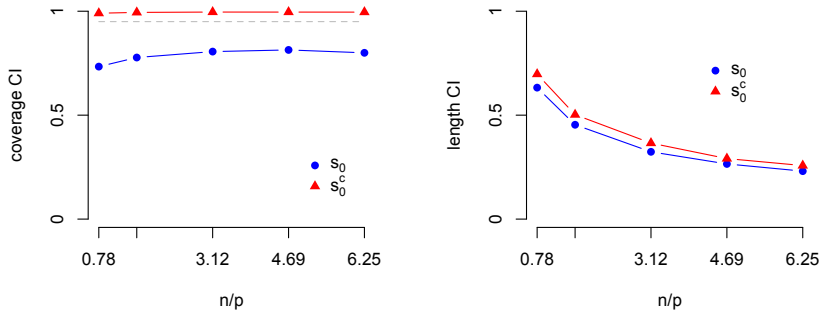


FIG 4. *The coverage (left panel) and length (right panel) of the confidence intervals, averaged over all intervals of the edge weights for different ratios n/p. The confidence intervals for the s_0 nonzero coefficients (blue circles) are separately depicted from the s_0^c zero coefficients (red triangles).*

Each row $\hat{\Theta}_{G,s}$ is $\hat{\beta}_{dL,si}/\hat{\tau}_s^2$ if $s \neq i$ and $1/\hat{\tau}_s^2$ if $s = i$. We have that $1/\hat{\tau}_s^2 = O_p(1)$ (see van de Geer et al., 2014, lemma 5.3), and so the ratio remains bounded. The desparsified lasso estimate is

$$\hat{\beta}_{dL,s} = \hat{\beta}_{L,s} - \hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}[X_{\setminus s}(\hat{\beta}_{L,s} - \beta_s)]/n + \hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}\varepsilon_s/n$$

So we must have the bound for the regular lasso and the bounds on $\hat{\Theta}_s$ and $X_{\setminus s}$. It follows from Theorem 1 of Raskutti, Wainwright and Yu (2010) that for the lasso estimate $\hat{\beta}_{L,s}$, for all $s \in V$, the ℓ_2 norm is $||\hat{\beta}_{L,s} - \beta_s||_2 = O_p(\sqrt{s_0 \log(p-1)/n})$ and $||X_{\setminus s}(\hat{\beta}_{L,s} - \beta_s)||_2/n = O_p(\sqrt{s_0 \log(p-1)/n})$ (see

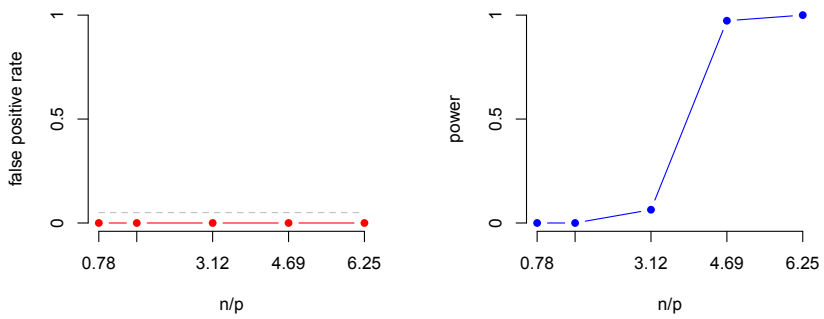


FIG 5. *The false positive rate (left) and power (right) of the test statistic* $T_{n,p}(X_1, X_2)$ *for different numbers of observations per parameter* n/p*, with* $p = 64$ *and* $n = 50$, 100, 200, 300, *and* 400.

also Section 5.4 in van de Geer et al., 2014). Additionally, for the ith row of $\hat{\Theta}_s$ we have $||\hat{\Theta}_{s,i}||_1 = O_p(\sqrt{s_0})$ (van de Geer et al., 2014), from which it follows that $||\hat{\Theta}_{s,i}||_2 = \sqrt{s_0}O_p(\sqrt{s_0})$. We then have that

$$\begin{aligned}||\hat{\beta}_{dL,s} - \beta_s||_2 &= ||\hat{\beta}_{L,s} - \beta_s - \hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}[X_{\setminus s}(\hat{\beta}_{L,s} - \beta_s)]/n + \hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}\varepsilon_s/n||_2 \\ &\leq ||\hat{\beta}_{L,s} - \beta_s||_2 + ||\hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}[X_{\setminus s}(\hat{\beta}_{L,s} - \beta_s)]/n||_2 + ||\hat{\Theta}_s X_{\setminus s}^{\mathsf{T}}\varepsilon_s/n||_2 \\ &\leq O_p\left(\sqrt{\frac{s_0\log(p-1)}{n}}\right) + O_p\left(\sqrt{\frac{s_0^3\log(p-1)}{n}}\right) + O_p\left(\frac{\sqrt{s_0}}{n}\right)\end{aligned}$$

The result follows if the sparsity s_0 for each $s \in V$ is $o(\sqrt[3]{n}/\log(p-1))$. □

Proof (Theorem 3) We first use the bound in Corollary 2 to determine the distribution of $T_{n,p}(X_1, X_2)$, and then generalise to multiple groups. We shall determine the expectation and variance of $T_{n,p}$ and then invoke the Hájek-Sidak central limit theorem to prove normality.

Rewriting the Frobenius norm by introducing the true values $\beta_{j,s}$ we obtain

$$\begin{aligned}||\hat{B}_1 - \hat{B}_2||_F^2 = \sum_{s\in V} &||\beta_{1,s} - \beta_{2,s}||_2^2 + ||\hat{\beta}_{1,dL,s} - \beta_{1,s}||_2^2 + ||\hat{\beta}_{2,dL,s} - \beta_{2,s}||_2^2 \\ &+ 2(\beta_{1,s} - \beta_{2,s})^{\mathsf{T}}(\hat{\beta}_{1,dL,s} - \beta_{1,s}) - 2(\beta_{1,s} - \beta_{2,s})^{\mathsf{T}}(\hat{\beta}_{2,dL,s} - \beta_{2,s}) \\ &+ (\hat{\beta}_{1,dL,s} - \beta_{1,s})^{\mathsf{T}}(\hat{\beta}_{2,dL,s} - \beta_{2,s})\end{aligned}$$

By the Cauchy-Schwartz inequality the first two cross products converge in probability to 0 when the conditional expectation is taken. Fix some $\epsilon > 0$ and

a finite $t \in \mathbb{R}^{p-1}$. Then

$$\begin{aligned}\mathbb{E}[|t^\mathsf{T}(\hat{\beta}_{2,dL,s} - \beta_{2,s})|]^2 &= \mathbb{E}[|t^\mathsf{T}(\hat{\beta}_{2,dL,s} - \beta_{2,s})|\mathbb{1}\{|t^\mathsf{T}(\hat{\beta}_{2,dL,s} - \beta_{2,s})| > \epsilon\}]^2 \\ &\leq \mathbb{E}[|t^\mathsf{T}(\hat{\beta}_{2,dL,s} - \beta_{2,s})|^2] \max_{1\leq i\leq p-1} t_i \mathbb{P}[||\hat{\beta}_{2,dL,s} - \beta_{2,s}||_1 > \epsilon]\end{aligned}$$

where the expectation and probability are conditional on $X_{\setminus s}$. Since $\mathbb{P}[||\hat{\beta}_{2,dL,s} - \beta_{2,s}||_1] > \epsilon]$ converges to 0 (van de Geer et al., 2014, and by Corollary 2), we have the result. The last cross product is 0 because the groups are assumed independent; applying the previous argument results in convergence to 0. Thus we have (10). And so we can easily obtain the mean (11) and variance (12).

Here we use the Hájek-Sidak central limit theorem to prove normality under H_0. For group j we write

$$||\hat{\Theta}_{j,s} X_{j,\setminus s}^\mathsf{T} \varepsilon_{j,s}||_2^2/n_j^2 - \mathbb{E}[||\hat{\Theta}_{j,s} X_{j,\setminus s}^\mathsf{T} \varepsilon_{j,s}||_2^2]/n_j^2 = \frac{1}{n_j^2} \sum_{i=1}^{n_j} \sum_{k=1}^{n_j} c_{j,s,ik}(\varepsilon_{j,s,i}\varepsilon_{j,s,k} - \sigma^2\delta_{ik})$$

where $c_{j,s,ik} = x_{j,\setminus s,i}^\mathsf{T} \hat{\Theta}_{j,s}^\mathsf{T} \hat{\Theta}_{j,s} x_{j,\setminus s,k}$ and $x_{j,\setminus s,i}$ is the ith column of $X_{j,\setminus s}$. It follows that $c_{j,s}^\mathsf{T} c_{j,s}/n_j^2 = \text{tr}(\hat{\Omega}_{j,s})^2$. Since the $\epsilon_{j,s} = \varepsilon_{j,s,i}\varepsilon_{j,s,k} - \sigma^2\delta_{ik}$ are independent, zero mean random variables with finite variance by assumption, and the variance of $c_{j,s}^\mathsf{T}\epsilon_{j,s}$ is $\sigma_j^4 c_{j,s}^\mathsf{T} c_{j,s}$, which is $n_j^2\sigma_j^4 \text{tr}(\hat{\Omega}_{j,s})^2$, the Hájek-Sidak central limit theorem (e.g., DasGupta, 2008, chap. 5) gives

$$\frac{\sum_{i=1}^{n_j}\sum_{k=1}^{n_j} c_{j,s,ik}(\varepsilon_{j,s,i}\varepsilon_{j,s,k} - \sigma^2\delta_{ik})}{\sigma_j^2\sqrt{\sum_{i,k=1}^{n} c_{j,s,ik}^2}} \xrightarrow{d} N(0,1)$$

Convergence holds if

$$\frac{\max_{ik} c_{j,s,ik}}{c_{j,s}^\mathsf{T} c_{j,s}} \to 0$$

as n increases. This condition holds because we assumed that the smallest eigenvalue of Σ is $\delta_{\min} > 0$ such that $1/\delta_{\min} = O(1)$ (Assumption 2), $||\hat{\Theta}_{j,s} - \Theta_{j,s}||_\infty = o_p(1)$ (Theorem 1), and the $X_{j,\setminus s}$ are by definition $O_p(1)$.

Scaling properly, we obtain $c_s^\mathsf{T} = (\phi_1^2 c_{1,s}^\mathsf{T}, \phi_2^2 c_{2,s}^\mathsf{T})$ with $\phi_j = \sigma_j/(\sqrt{n_j}\sigma)$ and $\epsilon_s^\mathsf{T} = (\epsilon_{1,s}^\mathsf{T}, \epsilon_{2,s}^\mathsf{T})$. We can then write the statistic that tests whether the groups differ as

$$T_{n,p}(X_1, X_2) = \frac{\sum_{s\in V} ||\beta_{1,s} - \beta_{2,s}||_2^2 + \sum_{s\in V} c_s^\mathsf{T}\epsilon_s}{\hat{\sigma}^2\sqrt{\sum_{s\in V} c_s^\mathsf{T} c_s}} + o_p(1)$$

We established that $c_{j,s}^\mathsf{T}\epsilon_{j,s}/n_j$ is $N(0, \sigma_j^4 \text{tr}(\hat{\Omega}_{j,s})^2)$, from which it follows that under H_0 : $(\beta_{1,s} = \beta_{2,s}, s \in V)$ the statistic $T_{n,p}$ converges to $N(0,1)$. The extension to q groups results in $q(q-1)/2$ comparisons, and hence, scaling $\sum_{i<j} T_{n,p}(X_i, X_j)$ for all comparisons with $1/\sqrt{q(q-1)/2}$ gives the result. □

7. Discussion

We showed that the desparsified lasso can be used for Gaussian selection of edges in a graph with with guarantees similar to those in van de Geer et al. (2014); Javanmard and Montanari (2014) if the nodewise sparsity is of order $\sqrt[3]{n}/\log(p-1)$. In simulations we showed that the recovery is reasonable to good even for few observations. Additionally, we showed that a high-dimensional test can be devised that determines whether graphs are different. The test has the advantages that it does not require degrees of freedom nor does it require cumbersome inversion of a large-scale parameter covariance matrix, as in Wald (type) statistics. A downside of the test is that the test is conservative when the ratio n/p is below about 4.5. For higher ratios the test has small false positive rate and a power of nearly 1 to detect differences if there indeed are differences between graphs.

Appendix

Combining shrinkage and desparsified lasso

Here we show that the shrinkage estimator is also suitable to use as an approximate inverse to $\hat{\Sigma}_{\setminus s,\setminus s}$. The shrinkage estimator of Θ_s is based on a linear combination of the estimate $\hat{\Sigma}_{\setminus s,\setminus s} = X_{\setminus s}^{\mathsf{T}} X_{\setminus j}/n$ and the identity matrix I. Ledoit and Wolf (2004) showed that using estimates of optimal shrinkage weights results in an estimator with minimal expected quadratic loss. Specifically, a shrinkage estimator using weight $0 \leq \rho \leq 1$ is $\hat{\Sigma}_{\setminus s,\setminus s,\rho} = \rho\mu I + (1-\rho)\hat{\Sigma}_{\setminus s,\setminus s}$. Ledoit and Wolf (2004) show that the estimates of the parameter μ is $O_p(1)$. If we furthermore assume that the eigenvalues of $\hat{\Sigma}_{\setminus s,\setminus s}$ are bounded (which is true if p/n converges to a limit (Yin, Bai and Krishnaiah, 1988)), then we obtain the following result.

Lemma 6 Let $\hat{\Sigma}_{\setminus s,\setminus s} = X_{\setminus s}^{\mathsf{T}} X_{\setminus s}/n$ be the covariance matrix and $\hat{\Sigma}_{\setminus s,\setminus s,\rho}$ is the shrinkage estimate. If the eigenvalues $(\delta_s, s \in V\setminus s)$ are bounded, the sparsity Assumption 3 $s_0 = o(\sqrt[3]{n}/\log(p-1))$ holds, Assumptions 1 and 2 hold, and the parameter ρ is chosen such that $\rho = o(\sqrt{\log(p-1)}/[\sqrt{n} + \sqrt{\log(p-1)}])$, then in the desparsified estimate (6) the error is $||\Delta_s||_\infty = \sqrt{n}||(\hat{\Theta}_{s,\rho}\hat{\Sigma}_{\setminus s,\setminus s} - I_{p-1})(\hat{\beta}_{L,s} - \beta_s)||_\infty = o_p(1)$.

Proof By Hölder's inequality we have that

$$||\Delta_s||_\infty \leq \sqrt{n}||(\hat{\Theta}_{s,\rho}\hat{\Sigma}_{\setminus s,\setminus s} - I_{p-1})||_F||\hat{\beta}_{L,s} - \beta_s||_2$$

We have for the lasso that $||\hat{\beta}_{L,s} - \beta_s||_2 = O_p(\sqrt{s_0 \log(p-1)/n})$ (see Corollary 2). We therefore only need a bound on the first term on the right hand side.

Let the eigenvalues of $\hat{\Sigma}_{\backslash s,\backslash s}$ be $(\delta_s, s \in V \backslash s)$ which are in the diagonal matrix D, and the corresponding eigenvectors in the orthonormal matrix U. Then the shrinkage estimator of $\hat{\Sigma}_{\backslash s,\backslash s}$ can be written as $\hat{\Sigma}_{\backslash s,\backslash s,\rho} = U(\rho\mu I + (1-\rho)D)U'$, and its inverse $\hat{\Sigma}^{-1}_{\backslash s,\backslash s,\rho} = \hat{\Theta}_{s,\rho}$

$$\hat{\Theta}_{s,\rho} = \sum_{s \in V} u_s u_s^{\mathsf{T}} \frac{\delta_s}{\rho\mu + (1-\rho)\delta_s}$$

The Frobenius norm is invariant to orthonormal transformations, and so we can consider

$$||\hat{\Theta}_{s,\rho}\hat{\Sigma}_{\backslash s,\backslash s} - I_{p-1}||_F \leq \max_{s \in V} \left| \frac{\delta_s}{\rho\mu + (1-\rho)\delta_s} - 1 \right|$$

It is easily checked that if $\rho = o(\sqrt{\log(p-1)}/[\sqrt{n} + \sqrt{\log(p-1)}])$, the estimate of μ is $O_p(1)$ (Ledoit and Wolf, 2004), and the largest eigenvalue is $O_p(1)$, then $||\hat{\Theta}_{s,\rho}\hat{\Sigma}_{\backslash s,\backslash s} - I_{p-1}||_F \leq O_p(\sqrt{\log(p-1)/n})$. Putting the two bounds together gives the result. □

References

Andrews, D. W. K. (1987). Asymptotic results for generalized Wald tests. *Econometric Theory* **3** 348-358.

Bai, Z., Jiang, D., Yao, J.-F. and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics* 3822–3840.

Bilodeau, M. and Brenner, D. (1999). *Theory of multivariate statistics*. New York: Springer-Verlag.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer-Verlag, New York.

Dufour, J.-M. and Valéry, P. (2011). Wald-type tests when rank conditions fail: a smooth regularization approach Technical Report, Working paper.

Ferguson, T. S. (1996). *A course in large sample theory*. Chapman and Hall, Bury st Edmunds.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression Technical Report, arXiv:1306.317.

Jöreskog, K. G. and Sörbom, D. (1989). *LISREL 7: A guide to the program and applications*. Chicago: SPSS Inc.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics* 1081–1102.

Ledoit, O. and Wolf, M. (2004). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis* **88** 365-411.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R. et al. (2014). A significance test for the lasso. *The Annals of Statistics* **42** 413–468.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34** 1436-1462.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104**.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. New York: John Wiley & Sons.

Nickl, R., van de Geer, S. et al. (2013). Confidence sets in sparse regression. *The Annals of Statistics* **41** 2852–2876.

Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* **100** 2065–2082.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11** 2241–2259.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis* **51** 6535–6542.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society* **35** 251–272.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.

Supekar, K., Musen, M. and Menon, V. (2009). Development of Large-Scale Functional Brain Networks in Children. *PloS Biology* **7** e1000157.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A. and Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific reports* **4**.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166-1202.

Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1988). On the limit of the

largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields* **78** 509-521.

Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). The Huge Package for High-dimensional Undirected Graph Estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062.