A U-statistic Approach to Hypothesis Testing for Structure Discovery in Undirected Graphical Models

Wacha Bounliphone, Matthew Blaschko

A U-statistic Approach to Hypothesis Testing for Structure Discovery in Undirected Graphical Models

Wacha Bounliphone* **and Matthew B. Blaschko†**

Inria Saclay, Galen Team
CentraleSupélec, L2S & CVN, Université Paris-Saclay
Grande Voie des Vignes
92295 Châtenay-Malabry, France

Center for Processing Speech & Images
Departement Elektrotechniek, KU Leuven
Kasteelpark Arenberg 10
3001 Leuven, Belgium
e-mail: wacha.bounliphone@centralesupelec.fr; matthew.blaschko@esat.kuleuven.be

Abstract: Structure discovery in graphical models is the determination of the topology of a graph that encodes conditional independence properties of the joint distribution of all variables in the model. For some class of probability distributions, an edge between two variables is present if and only if the corresponding entry in the precision matrix is non-zero. For a finite sample estimate of the precision matrix, entries close to zero may be due to low sample effects, or due to an actual association between variables; these two cases are not readily distinguishable. Many related works on this topic consider potentially restrictive distributional or sparsity assumptions that may not apply to a data sample of interest, and direct estimation of the uncertainty of an estimate of the precision matrix for general distributions remains challenging. Consequently, we make use of results for U-statistics and apply them to the covariance matrix. By probabilistically bounding the distortion of the covariance matrix, we can apply Weyl's theorem to bound the distortion of the precision matrix, yielding a conservative, but sound test threshold for a much wider class of distributions than considered in previous works. The resulting test enables one to answer with statistical significance whether an edge is present in the graph, and convergence results are known for a wide range of distributions. The computational complexities is linear in the sample size enabling the application of the test to large data samples for which computation time becomes a limiting factor. We experimentally validate the correctness and scalability of the test on multivariate distributions for which the distributional assumptions of competing tests result in underestimates of the false positive ratio. By contrast, the proposed test remains sound, promising to be a useful tool for hypothesis testing for diverse real-world problems. Source code for the tests is available for download from https://github.com/wbounliphone/Ustatistics_Approach_For_SD.

Keywords and phrases: significance hypothesis testing, covariance matrix, precision matrix, structure discovery, U-statistics estimator.

*WB is supported in part by a CentraleSupélec Fellowship and ERC Grant 259112.
†This work is supported in part by Internal Funds KU Leuven and FP7-MC-CIG 334380. We thank Jonas Peters for helpful discussions and comments on an early draft of this work.

Contents

1. Introduction

Graphical models are powerful tools for analyzing relationships between a set of random variables, so that key conditional independence properties can be read from a graph. Learning the structure of an underlying graphical model is of fundamental importance and has applications in a large number of domains - e.g. analysis of fMRI brain connectivity, analysis of genes associated with complex human diseases, or analysis of interactions in social networks. In many contemporary applications, a large, effectively unlimited stream of raw data with unknown multivariate distribution is to be analyzed. In such scenarios, computation becomes a fundamental limit and methods that can estimate properties of graphical models from very general distributions with computation linear in the number of observations become necessary. We address this problem setting in this paper by devising a probabilistic bound on the entries of the precision matrix for highly general distributions that decreases in the sample size as $\mathcal{O}(n^{-1/2})$, while maintaining linear time computation. This bound can

then be used to construct a hypothesis test for a graphical model structure, or for upper and lower bounds on the effect between two variates.

We can divide graphical models in two types, namely directed graphical models, e.g. Bayesian networks (Pearl, 2014; Jensen, 1996; Neapolitan, 2004) or undirected graphical models, e.g. Gaussian graphical models (Whittaker, 2009; Lauritzen, 1996; Speed and Kiiveri, 1986). Here, we focus on undirected graphical models to exhibit the conditional dependence structure in multivariate distributions.

Hypothesis testing with statistical measures of dependence is a relatively well developed field with a number of general results. Classical tests such as Spearman's ρ (Spearman, 1904), Kendall's τ (Kendall, 1938), Rényi's α (Rényi, 1961) and Tsallis' α (Tsallis, 1988) are widely applied. Recently, for multivariate nonlinear dependencies, novel statistical tests were introduced and some prominent examples include the kernel mutual information (Gretton et al., 2003), the generalized variance and kernel canonical correlation analysis (Bach and Jordan, 2003), the Hilbert-Schmidt independence criterion (Gretton et al., 2005), the distance based correlation (Székely et al., 2007) and rankings (Heller et al., 2012). Testing the conditional dependence is even more challenging, and only few dependence measures have been generalized to the conditional case (Fukumizu et al., 2007, 2009; Zhang et al., 2011). We note that their work requires the estimate of a regularization parameter with appropriate asymptotic decrease to estimate the distribution of the test statistic under the null hypothesis, as well as for kernel selection, and has quadratic space usage rendering it inapplicable to very large data sets. Futhermore, Roverato and Whittaker (1996) provided an asymptotic distribution for the inverse covariance which is Gaussian and this required the computation of the Isserlis matrix of the inverse of the covariance matrix. These results, however, do not directly extend to the test that we analyze here: that of independence between two variables *conditioned* on all the others:

$$X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}. \tag{1.1}$$

In the case of multivariate Gaussian distribution, the non-zero entry in the inverse of the covariance matrix can be shown to correspond to the underlying structure of the graphical model (Dempster, 1972). This observation has motivated a range of structure discovery techniques in high-dimensional settings, where $n < p$ (see Table 1 for notation). Estimation of such high-dimensional models has been the focus on recent research (Schäfer and Strimmer, 2005; Li and Gui, 2006; Meinshausen and Bühlmann, 2006; Banerjee et al., 2008; Friedman et al., 2008; Ravikumar et al., 2011) where methods impose a sparsity constraint on the entries of the inverse covariance matrix. The consequence of this attractive method to estimate the inverse of the sparse covariance matrix has been the development of diverse statistical hypothesis tests (G'Sell et al., 2013; Lockhart et al., 2014; Janková and van de Geer, 2015). Each of these methods explicitly assumes that the data distribution is multivariate Gaussian.

By contrast, we instead focus in this paper on designing a test for the $n > p$ case, and in particular ensure that the test has computational complexity *linear* in n, while making minimal distributional assumptions. These assumptions are: (i) that the covariance matrix exists and an unbiased estimate converges to this matrix (cf. Theorem 4), and (ii) that the eigenvector-eigenvalue product converges at most at the same asymptotic rate as the convergence of the eigenvalues (cf. Lemma 2 and Xia et al. (2013)).

In the case of non-Gaussian graphical models, several techniques focus on the existence of a relationship between conditional independence and the structure of the inverse covariance matrix. Loh and Wainwright (2013) have established theoretical results by extending a number of interesting links between covariance matrices and the graphical model in the case of discrete random variables and particularly for tree-structured graphs.

While there exist many convenient methods using Gaussian multivariate distributions or discrete variables, other distributions pose new challenges in statistical modeling. Consequently, we develop a statistically and computationally efficient framework for hypothesis testing of whether an entry of the precision matrix is non-zero based on a data sample from the joint distribution P_X. The proposed test not only has asymptotic guarantees, but is sound for all finite sample sizes without the need to set a regularization parameter or perform a computationally expensive bootstrap procedure.

In this paper, we have taken the approach of precisely modeling the joint distribution of the covariance matrix, and using this distribution to probabilistically bound the distortion of the covariance matrix. The joint distribution of the entries of the covariance matrix is asymptotically Gaussian with known parameters due to the theory of U-statistics (Serfling, 2009; Lehmann, 1999; Hoeffding, 1948; Lee, 1990). We are then able to make use of Weyl's theorem (Weyl, 1912) to upper bound the distortion of the precision matrix as a function of the distortion of the covariance matrix, which yields an upper bound on the test threshold at a given significance level. We derive two upper bounds on the test threshold, one of which is strictly tighter than the other, with computational complexities $\mathcal{O}(np^2 + p^3)$ and $\mathcal{O}(np^4)$, respectively, where n is the sample size and p is the number of variables. We also present a simulation study illustrating analytically and experimentally that both of these thresholds are sound for a substantially more general set of distributions compared with competing tests in the literature and decrease as $\mathcal{O}(n^{-1/2})$.

2. Preliminary definitions

In this section, we give a brief background of undirected graphical models and testing conditional independence (section 2.1 and section 2.2) and a basic de-

TABLE 1
Notation Table

Notation	Description
$\mathrm{G} = (\mathrm{V}, \mathrm{E})$	Graph G, where V is a finite set of vertices with $\lvert V\rvert = d$, $\mathrm{E} \subseteq \mathrm{V} \times \mathrm{V}$ is a subset of ordered pairs of distinct vertices (i, j);
X	$\mathrm{X} = \{\mathrm{X}_1, ..., \mathrm{X}_p\}$ is a set of random variables of dimension p with sample size n;
Σ	Covariance matrix of X;
$\hat{\Sigma}$	Unbiased estimator of the covariance matrix of X estimated from n samples;
Θ	Precision matrix of Σ;
$\hat{\Theta}$	Empirical estimate of the precision matrix;
$\overline{X}$ and $\overline{XY}$	$\mathrm{E}[X]$ and $\mathrm{E}[XY]$;
$(\mathcal{T}_{ij}, \hat{\Theta}_{ij}, \delta)$	The statistical test $\mathcal{T}_{ij}$ with statistic $\hat{\Theta}_{ij}$ at a significance level δ;
t	The threshold of the test statistic;
$U(A)$	Function returning the upper triangular part and diagonal of a matrix A

scription of the U-statistic estimator for the covariance matrix (section 2.3).

2.1. Undirected Graphical Models

Graphical models blend probability theory and graph theory together. They are powerful tools for analyzing relationships between a large number of random variables (Whittaker, 2009; Lauritzen, 1996; Koller and Friedman, 2009). A *graph* is set of vertices $\mathrm{V} = \{1, ...p\}$ and a set of edges $\mathrm{E(G)} \subseteq \mathrm{V} \times \mathrm{V}$. We study undirected graphical models (also known as Markov random fields).

Undirected Graphical model An *undirected graphical model* is a joint probability distribution, P_X, defined on an undirected graph G, where the vertices V in the graph index a collection of random variables $\mathrm{X} = \{\mathrm{X}_1, ..., \mathrm{X}_p\}$ and the edges encode conditional independence relationships among random variables

$$P_X \propto \prod_{c \in \mathcal{C}} \Psi_c(X_c) \tag{2.1}$$

where $\mathcal{C}$ is the set of maximal cliques in the graph and $\{\Psi_c\}_{c \in \mathcal{C}}$ are non-negative potential functions.

2.2. Testing conditional independence in undirected graphical models

Conditional independence (CI) is an important concept in statistics, artificial intelligence, and related fields (Dawid, 1979). A common measure for the testing of independence of two variables conditioned on a third variable is the *partial correlation* $\rho_{XY.Z}$. With the assumption that all variables are multivariate Gaussian, the partial correlation is zero if and only if X is conditionally independent from Y given Z

$$H_0 : \rho_{XY.Z} = 0 \qquad \text{vs} \qquad H_1 : \rho_{XY.Z} \neq 0. \tag{2.2}$$

The distribution of the sample partial correlation was described by Fisher (Fisher, 1924) and we would reject H_0 if the absolute value of the test statistic exceeded the critical value from the Student table evaluated at $\delta/2$. The computational complexity of the partial correlation is $\mathcal{O}(np^2+p^3)$ which simplifies to $\mathcal{O}(np^2)$ as $n \geq p$. However, as mentioned in Kendall (1946, Chap. 26 & 27), this hypothesis test makes a strong assumption that the data are Gaussian distributed, and in particular that the fourth-order moment is equal to 0.

Furthermore, tests of conditional independence can be made without any assumption of normality in the distribution, using for instance the permutation distribution of $\rho_{XY.Z}$ or bootstrap techniques, but this becomes too computationally expensive in practice when n tends to be large.

2.3. A U-statistic Estimator of the Cross-Covariance

Most of the materials in this subsection can be found in Hoeffding (1948), Serfling (2009, Chap. 5), Lehmann (1999, Chap. 6) and Lee (1990). Suppose we have a sample $\mathrm{X} = \{(X_{i_1}, ...X_{i_p})\}_{1 \leq i \leq n}$ of size n drawn i.i.d. from a distribution P_X. A U-statistic concerns an unbiased estimator of a parameter θ of P_X using X. Suppose there is some function $h(X_1, ..., X_q)$ which is an unbiased estimator of $\theta = \mathrm{E}[h(X_1, ...X_q)]$, h is called a kernel of order $q \leq p$ of the estimator. When we have a sample $\mathrm{X} = \{X_{i_1}, ...X_{i_q})\}_{1 \leq i \leq n}$ of size n larger than p, we can then construct a U-statistic in the following way.

Definition 1. *(U-statistic) Given a kernel h of order q and a sample $\mathrm{X} = \{X_{i_1}, ...X_{i_q})\}_{1 \leq i \leq n}$ of size n larger than p, the corresponding U-statistic for estimation of θ is obtained by the following*

$$\hat{U} := \frac{1}{(n)_q} \sum_{i_q^n} h(X_{i_1}, ..., X_{i_q}) \tag{2.3}$$

where the summation ranges over q indices drawn without replacement from $(1, ..., n)$ and $(n)_q$ is the Pochhammer symbol $(n)_q := \dfrac{n!}{(n-q)!}$.

Definition 2. *(U-statistic estimator of the covariance) Let* $u_i = (X_i, Y_i)^T$ *be ordered pairs of samples* $1 \leq i \leq p$*. Consider* $\Sigma = \mathrm{Cov}(X, Y)$*, the covariance functional between* X *and* Y *and* h*, the kernel of order 2 for the functional* Σ *such that*

$$h(u_1, u_2) = \frac{1}{2}(X_1 - X_2)(Y_1 - Y_2). \tag{2.4}$$

The corresponding U-statistic estimator of the covariance Σ *is*

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)(Y_i - Y_j) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{2.5}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$*.* $\hat{\Sigma}$ *can be computed in linear time.*

3. Structure Discovery in Undirected Graphical Models

In this section, we will use the U-statistic estimator of the covariance matrix to define a hypothesis test for discovering the structure of graphical models. We show that this estimator can be computed in time linear in the number of samples and study its concentration distribution. We will denote the covariance matrix by Σ with its unbiased estimator $\hat{\Sigma}$ using Definition 2, and $\Theta = \Sigma^{-1}$ for the precision matrix, with $\hat{\Theta}$ its empirical estimate.

3.1. Discovery based on a U-statistic estimator

As the distribution of $\hat{\Theta}$ under the null hypothesis is unknown in general, we focus here on U-statistic estimates of $\hat{\Sigma}$ and its asymptotic normal distribution to calculate conservative bounds on the threshold for our hypothesis test. We therefore develop the full covariance between the elements of $\hat{\Sigma}$, which we denote $\mathrm{Cov}(\hat{\Sigma}) \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}$. The size of $\mathrm{Cov}(\hat{\Sigma})$ is due to the symmetry of $\hat{\Sigma}$.

Theorem 1. *(Joint asymptotic normality distribution of the covariance matrix) For all* (i, j, k, l) *range over each of the* p *variates in a covariance matrix* $\hat{\Sigma}$*, if* $\mathrm{Var}(\hat{\Sigma}_{ij}) > 0$ *and* $\mathrm{Var}(\hat{\Sigma}_{kl}) > 0$*, then*

$$n^{\frac{1}{2}}\begin{pmatrix}\hat{\Sigma}_{ij} - \Sigma_{ij} \\ \hat{\Sigma}_{kl} - \Sigma_{kl}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\mathrm{Var}(\hat{\Sigma}_{ij}) & \mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) \\ \mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) & \mathrm{Var}(\hat{\Sigma}_{kl})\end{pmatrix}\right). \tag{3.1}$$

Theorem 2. *(Covariance of the U-statistic for the covariance matrix)*

We note respectively h *and* g *the corresponding kernel of order 2 for the two*

unbiased estimates $\hat{\Sigma}_{ij}$ *and* $\hat{\Sigma}_{kl}$*, where*

$$h(u_1, u_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{j_1} - X_{j_2}\right), \text{with } u_r = (X_{i_r}, X_{j_r})^T \tag{3.2}$$

$$g(v_1, v_2) = \frac{1}{2}\left(X_{k_1} - X_{k_2}\right)\left(X_{l_1} - X_{l_2}\right), \text{with } v_r = (X_{k_r}, X_{l_r})^T. \tag{3.3}$$

The low variance, unbiased estimates of the covariance between two U*-statistics estimates* $\hat{\Sigma}_{ij}$ *and* $\hat{\Sigma}_{kl}$*, where* (i, j, k, l) *range over each of the* p *variates in a covariance matrix* $\hat{\Sigma}$ *is*

$$\mathrm{Cov}(\hat{\Sigma}) := \mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) = \binom{n}{2}^{-1}(2(n-2)\zeta_1) + \mathcal{O}(n^{-2}) \tag{3.4}$$

where $\zeta_1 = \mathrm{Cov}\left(\mathrm{E}_{u_2}[h(u_1, u_2)], \mathrm{E}_{v_2}[g(v_1, v_2)]\right)$.

Proof. Eq. (3.4) is constructed with the definition of Covariance of a U-statistic as given by Hoeffding (1948). □

Theorem 3. *There are seven exhaustive cases which can be used to estimate Eq. (3.4) for all* $1 \leq i, j, k, l \leq p$ *through simple variable substitution. Each of these cases has computation linear in* n*.*

Case 1: $i \neq j, k, l;\ j \neq k, l;\ k \neq l$

$$\begin{aligned}\zeta_1 = \frac{1}{4}\Big\{&\overline{X_iX_jX_kX_l} - \overline{X_i}\ \overline{X_jX_kX_l} - \overline{X_j}\ \overline{X_iX_kX_l} \\ &- \overline{X_k}\ \overline{X_iX_jX_l} + \overline{X_i}\ \overline{X_k}\ \overline{X_jX_l} + \overline{X_j}\ \overline{X_k}\ \overline{X_iX_l} \\ &- \overline{X_iX_jX_k}\ \overline{X_l} + \overline{X_i}\ \overline{X_l}\ \overline{X_jX_k} + \overline{X_j}\ \overline{X_l}\ \overline{X_iX_k} \\ &- \left(\overline{X_iX_j} - 2\ \overline{X_i}\ \overline{X_j}\right)\left(\overline{X_kX_l} - 2\ \overline{X_k}\ \overline{X_l}\right)\Big\}\end{aligned} \tag{3.5}$$

Case 2: $i = j;\ j \neq k, l;\ k = l$

$$\begin{aligned}\zeta_1 = \frac{1}{4}\Big\{&\overline{X_i^2X_k^2} - 2\ \overline{X_i}\ \overline{X_iX_k^2} - 2\ \overline{X_i^2X_{k_1}}\ \overline{X_k} + 4\overline{X_iX_k}\ \overline{X_i}\ \overline{X_k} \\ &- \left(\overline{X_i^2} - 2\ \overline{X_i}^2\right)\left(\overline{X_k^2} - 2\ \overline{X_k}^2\right)\Big\}\end{aligned} \tag{3.6}$$

Case 3: $i = j;\ j \neq k, l;\ k \neq l$

$$\begin{aligned}\zeta_1 = \frac{1}{4}\Big\{&\overline{X_i^2X_kX_l} - 2\ \overline{X_iX_kX_l}\ \overline{X_i} - \overline{X_i^2X_l}\ \overline{X_k} \\ &+ 2\ \overline{X_iX_l}\ \overline{X_i}\ \overline{X_k} - \overline{X_i^2X_{k_1}}\ \overline{X_l} + 2\ \overline{X_iX_k}\ \overline{X_i}\ \overline{X_l} \\ &- \left(\overline{X_i^2} - 2\ \overline{X_i}^2\right)\left(\overline{X_kX_l} - 2\ \overline{X_k}\ \overline{X_l}\right)\Big\}\end{aligned} \tag{3.7}$$

Case 4: $i = k;\ j \neq i, k, l;\ k \neq l$

$$\begin{aligned}\zeta_1 = \frac{1}{4}\Big\{ & \overline{X_{i_1}^2 X_{j_1} X_{l_1}} - \overline{X_i}\ \overline{X_{j_1} X_{i_1} X_{l_1}} - \overline{X_{i_1}^2 X_{l_1}}\ \overline{X_j} \\ & - \overline{X_{i_1} X_{j_1} X_{l_1}}\ \overline{X_i} + \overline{X_i}^2\ \overline{X_{j_1} X_{l_1}} + \overline{X_{i_1} X_{l_1}}\ \overline{X_j}\ \overline{X_i} \\ & - \overline{X_{i_1}^2 X_{j_1}}\ \overline{X_l} + \overline{X_i}\ \overline{X_{j_1} X_{i_1}}\ \overline{X_l} + \overline{X_{i_1}^2}\ \overline{X_j}\ \overline{X_l}] \\ & - \left(\overline{X_i X_j} - 2\ \overline{X_i}\ \overline{X_j}\right)\left(\overline{X_i X_l} - 2\ \overline{X_i}\ \overline{X_l}\right)\Big\} \end{aligned} \tag{3.8}$$

Case 5: $i = k;\ i \neq j;\ j = l;$

$$\begin{aligned}\zeta_1 = \frac{1}{4}\Big\{ & \overline{X_i^2 X_j^2} - 2\overline{X_i X_j^2}\ \overline{X_i} + \overline{X_i}^2\ \overline{X_j^2} - 2\overline{X_i^2 X_j}\ \overline{X_j} + 2\overline{X_i}\ \overline{X_j}\ \overline{X_j X_i} + \overline{X_i^2}\ \overline{X_j}^2 \\ & - \left(\overline{X_i X_j} - 2(\overline{X_i}\ \overline{X_j})\right)^2\Big\} \end{aligned} \tag{3.9}$$

Case 6: $i = j = k;\ i \neq l$

$$\begin{aligned}\zeta_1 = \frac{1}{4}\Big\{ & \overline{X_i^3 X_l} - 3\ \overline{X_i^2 X_l}\ \overline{X_i} + 2\ \overline{X_i X_l}\ \overline{X_i}^2 - \overline{X_i^3}\ \overline{X_l} + 2\ \overline{X_i^2}\ \overline{X_i}\ \overline{X_l} \\ & - \left(\overline{X_i^2} - 2\ \overline{X_i}^2\right)\left(\overline{X_i X_l} - 2\ \overline{X_i}\ \overline{X_l}\right)\Big\} \end{aligned} \tag{3.10}$$

Case 7: $i = j, k, l$

$$\zeta_1 = \frac{1}{4}\left\{\overline{X_i^4} - 4\overline{X_i^3}\ \overline{X_i} + 4\overline{X_i^2}\ \overline{X_i}^2 - \left(\overline{X_i^2} - 2\overline{X_i}^2\right)^2\right\} \tag{3.11}$$

Proof. A proof of Theorem 3 is given in Appendix A. □

We now have that an estimator of a covariance matrix has asymptotic joint Gaussian distribution of its entries. This may appear contrary to the fact that a covariance matrix lies in the positive definite cone as a Gaussian distribution has unbounded support. We show here that a Gaussian distribution does not contradict a positive definite covariance matrix by demonstrating concentration of the probability distribution in the positive definite cone.

Theorem 4. *(Concentration of probability) Let us assume that* X *has finite support* $[a, b]$ *with probability at least* $1 - \gamma$ *for some distribution dependent* $\gamma \geq 0$*, then for* $n > 1$ *and all* $\delta > 0$*, with probability at least* $(1 - \delta)(1 - \gamma)$ *for all* P_X

$$|\hat{\Sigma}_{ij} - \Sigma_{ij}| \leq (b - a)\sqrt{\log(\delta/2)/n} \quad \forall i, j. \tag{3.12}$$

Proof. The estimator $\widehat{\Sigma}$ of the covariance matrix Σ is a U-statistic of order 2, where each term is contained in $[a, b]$. By using the concentration inequality of

Hoeffding for U-statistics, we achieve

$$2\exp\left(-\frac{2(n/2)\varepsilon^2}{(b-a)^2}\right) = \delta \tag{3.13}$$

and obtain $\varepsilon = (b-a)\sqrt{\log(\delta/2)/n}$. □

If $(1-\delta)(1-\gamma)$ can approach 1 arbitrarily closely while the r.h.s. of Eq. (3.12) goes to zero, this concentration of probability will mean that once a sufficient data sample are observed, the maximum eigenvalue of $\mathrm{Cov}(\hat{\Sigma})$ will be much smaller than the smallest eigenvalue of Σ, and the distribution will be concentrated in the positive definite cone. An explicit bound on the concentration in the positive definite cone based on Weyl's theorem is employed in the following section to construct our test threshold.

3.2. Hypothesis Test using a U-statistic estimator for the covariance matrix

We now describe a statistical test for structure discovery in graphical models, based on the U-statistic estimator $\hat{\Sigma}$ of the covariance matrix. Given X a sample matrix of size $n \times p$ and for all $(i,j) \in \{1,...,p\}$, the statistical test $(\mathcal{T}_{ij}, \hat{\Theta}_{ij}, \delta) : (\mathrm{X}, i, j) \longmapsto \{0,1\}$, is used to distinguish between the following null hypothesis $H_0(i,j)$ and the two-sided alternative hypothesis $H_1(i,j)$:

$$H_0(i,j) : \Theta_{i,j} = 0 \qquad \text{vs} \qquad H_1(i,j) : \Theta_{i,j} \neq 0 \tag{3.14}$$

at a significance level δ. This is achieved by comparing the test statistic, $|\hat{\Theta}_{ij}|$ with a particular threshold t: if the threshold is exceeded, then the test rejects the null hypothesis. The acceptance region of the test is thus defined as any real number below the threshold.

In the following we will explain in Theorem 6 how the threshold is determined and show that it is a conservative bound. To prove Theorem 6, we make use of Lemmas 1 and 2.

Lemma 1. *With probability at least* $1-\delta$

$$\|\Sigma - \hat{\Sigma}\|_2 \leq \sqrt{2\lambda_{\max}}\Phi^{-1}\left(1-\delta/2\right) \tag{3.15}$$

where $\Phi(\cdot)$ *is the CDF of a standard normal distribution and* $\lambda_{\max}$ *is the largest eigenvalue of* $\mathrm{Cov}(\hat{\Sigma})$.

Proof. As $\hat{\Sigma}$ is a U-statistic, we have that $U(\hat{\Sigma})$, a vector containing its upper diagonal component (including the diagonal), is Gaussian distributed with covariance $\mathrm{Cov}(\hat{\Sigma})$ (cf. Thm 1, 2). Therefore, with probability at least $1-\delta$.

$$\|U(\Sigma) - U(\hat{\Sigma})\|_2 \leq \sqrt{\lambda_{\max}}\Phi^{-1}\left(1-\delta/2\right) \tag{3.16}$$

and furthermore

$$\|\Sigma - \hat{\Sigma}\|_F \leq \sqrt{2}\|U(\Sigma) - U(\hat{\Sigma})\|_2 \tag{3.17}$$

which combined with the fact that $\|\cdot\|_2 \leq \|\cdot\|_F$ yields the desired result. □

Corollary 1. *With probability with at least* $1-\delta$

$$\|\Sigma - \hat{\Sigma}\|_2 \leq \sqrt{2\,\mathrm{Tr}[\mathrm{Cov}(\hat{\Sigma})]}\Phi^{-1}\left(1-\delta/2\right) \tag{3.18}$$

Lemma 2. *(Bounding the deviation of the empirical precision matrix as a function of eigenvalues) Given* X *a set of random variables drawn from a distribution for which Eq.* (3.24 B) *converges at a rate* $\mathcal{O}(n^{-1/2})$ *with a precision matrix* Θ*, and an empirical estimate of the precision matrix* $\hat{\Theta}$ *corresponding to a covariance matrix* $\hat{\Sigma}$ *with eigenvalues* $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$*, then with high probability*

$$|\hat{\Theta}_{ij} - \Theta_{ij}| \leq \mu\sqrt{\sum_{k=1}^{p}\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2} \quad \forall i,j \in \{1,...,p\} \tag{3.19}$$

for a distribution dependent constant μ*.*

Proof. We denote respectively $\hat{\Sigma}$ the perturbed matrix of Σ, with $\alpha_1 \geq ... \geq \alpha_p$ the eigenvalues of Σ and $\hat{\alpha}_1 \geq ... \geq \hat{\alpha}_p$ the eigenvalues of an empirical estimate of the true covariance matrix $\hat{\Sigma}$, and $\hat{\Theta}$ the perturbed matrix of Θ. We then have that $|\hat{\Theta}_{ij} - \Theta_{ij}| \leq \|\hat{\Theta} - \Theta\|_F$ for all $i,j \in \{1,...,p\}$. We will use the property of the singular value decomposition that $\hat{\Sigma} = \hat{V}\hat{A}\hat{V}^T$, where $\hat{V}$ is an $n \times n$ unitary matrix and a diagonal matrix $\hat{A}$ with $\hat{A}_{ii} = \hat{\alpha}_i$ is the i-th eigenvalue of $\hat{\Sigma}$. Furthermore, we have that $\Sigma^{-1} = \Theta$ and the empirical estimate of Θ is $\hat{\Theta}$ such that $\hat{\Theta} = \hat{U}\hat{\Lambda}\hat{U}^T$ where $\hat{U}$ is an $n \times n$ unitary matrix and a diagonal matrix $\hat{\Lambda}$ with $\hat{\Lambda}_{ii} = 1/\hat{\alpha}_i$.

$$\begin{aligned}
\|\hat{\Theta} - \Theta\|_F^2 &= \mathrm{Tr}\left[(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)\right] && (3.20)\\
&= \mathrm{Tr}\left[\hat{\Theta}\hat{\Theta} + \Theta\Theta - 2\hat{\Theta}\Theta\right] && (3.21)\\
&= \mathrm{Tr}\left[\hat{\Lambda}\hat{\Lambda} + \Lambda\Lambda - 2\hat{U}\hat{\Lambda}\hat{U}^T U\Lambda U^T\right] && (3.22)\\
&= \mathrm{Tr}\left[\hat{\Lambda}\hat{\Lambda} + \Lambda\Lambda - 2\Lambda\hat{\Lambda}\right] + 2\,\mathrm{Tr}\left[\Lambda\hat{\Lambda} - \hat{U}\hat{\Lambda}\hat{U}^T U\Lambda U^T\right] && (3.23)\\
&= \underbrace{\sum_{k=1}^{p}\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2}_{3.24\text{ A}} + \underbrace{2\sum_{k=1}^{p}\frac{1}{\alpha_k\hat{\alpha}_k} - 2\,\mathrm{Tr}\left[\hat{U}\hat{\Lambda}\hat{U}^T U\Lambda U^T\right]}_{3.24\text{ B}} && (3.24)\\
&\leq \mu\left(\sum_{k=1}^{p}\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2\right) && (3.25)
\end{aligned}$$

The bound in Eq. (3.25) will hold with high probability, e.g. when the finite moment conditions of Xia et al. (2013) are satisfied, as Eq. (3.24) is then guaranteed to converge with rate $\mathcal{O}(n^{-1/2})$. □

We have now shown that we can compute a bound on the distortion purely from the eigenvalues of $\hat{\Sigma}$.

Theorem 5. *(Weyl's Theorem, Weyl (1912)) For two positive definite matrices* Σ *and* $\hat{\Sigma}$ *with corresponding eigenvalues* α_k *and* $\hat{\alpha}_k$*, respectively, if*

$$|\alpha_k - \hat{\alpha}_k| \leq \|\hat{\Sigma} - \Sigma\|_2 \leq \varepsilon \tag{3.26}$$

where $0 < \varepsilon < \alpha_k \ \forall k \in \{1, ..., p\}$*, then*

$$\alpha_k - \varepsilon \leq \hat{\alpha}_k \leq \alpha_k + \varepsilon \quad \forall k \in \{1, ..., p\}. \tag{3.27}$$

Theorem 6. *(Conservative threshold) For all* $(i, j) \in \{1, ..., p\}$*, the threshold* t *for testing* $H_0 : \Theta_{i,j} = 0$ *versus the alternative hypothesis* $H_1 : \Theta_{i,j} \neq 0$ *is given by* P *for a small probability* $\delta \in (0, 1)$ *such that*

$$\mathrm{P}\left(|\hat{\Theta}_{i,j}| > t | \Theta_{i,j} = 0\right) < \delta \tag{3.28}$$

where t *is a conservative threshold*

$$t = \mu \sqrt{\sum_{k=1}^{p} \left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2} \tag{3.29}$$

with $\hat{\alpha}_k$ *the* k*-th eigenvalue of the empirical covariance matrix* $\hat{\Sigma}$*,* μ *a distribution dependent constant satisfying the inequality (3.25), and* ε *is an error bound such that*

$$\varepsilon_{\text{Eig}} = \sqrt{2\lambda_{max}}\Phi\left(1 - \delta/2\right), \text{ or } \varepsilon_{\text{Trace}} = \sqrt{2\,\text{Tr}[\text{Cov}(\hat{\Sigma})]}\Phi\left(1 - \delta/2\right) \tag{3.30}$$

where λ_{max} *is the largest eigenvalue of* $\text{Cov}(\hat{\Sigma})$ *and* $\text{Tr}[\text{Cov}(\hat{\Sigma})]$ *is the trace of* $\text{Cov}(\hat{\Sigma})$*.*

Proof. We have shown that we can compute the distortion of $\hat{\Theta}$ purely from the eigenvalues of Σ and $\hat{\Sigma}$. Therefore, we use Weyl's theorem on the covariance matrix to get error bounds for the eigenvalues of Σ. Inequality (3.27) gives the following bounds for the eigenvalues of the precision matrix Θ

$$\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2 \leq \left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2 \quad \forall k \in \{1, ..., p\} \tag{3.31}$$

Combining Eq. (3.25) and (3.31) gives

$$\|\hat{\Theta} - \Theta\|_F \leq \mu \sqrt{\sum_{i=1}^{p} \left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2} \tag{3.32}$$

and

$$|\hat{\Theta}_{ij} - \Theta_{ij}| \leq \|\hat{\Theta} - \Theta\|_F. \tag{3.33}$$

□

Theorem 7. *For a fixed computational budget N less than the time required to process all data points, and for sufficiently large p, the trace bound decreases at the same asymptotic rate as the eigenvalue bound.*

Proof. We note that the bound in Corollary 1 is strictly larger than that of Lemma 1, but its computation $C_{\text{Trace}}(n,p) \asymp np^2$ as opposed to $C_{\text{Eig}}(n,p) \asymp np^4$, where $\asymp$ denotes that the function is asymptotically bounded above and below (Temlyakov, 2011). The number of samples processed is $n_{\text{Trace}}(N,p) \asymp N/p^2$ for the trace test and $n_{\text{Eig}}(N,p) \asymp N/p^4$ for the eigenvalue test.

For a full rank $p^2 - \binom{p}{2} \times p^2 - \binom{p}{2}$ p.s.d. matrix, the trace is $\mathcal{O}(p^2\lambda_{max})$. We have when the sample sizes are equal $\varepsilon_{\text{Trace}} \in \mathcal{O}(p\varepsilon_{\text{Eig}})$. Furthermore, Equation (3.29) is asymptotically linear in ε as ε approaches zero from the right, and $\varepsilon_{\text{Eig}} \in \mathcal{O}(\lambda(p)n^{-1/2})$, where $\lambda(p)$ gives the dependence of ε_{Eig} on the dimensionality of the data. Therefore, at a fixed computational budget the eigenvalue threshold is $\mathcal{O}(\lambda(p)n_{\text{Eig}}(n,p)^{-1/2}) = \mathcal{O}(\lambda(p)(Np^{-4})^{-1/2}) = \mathcal{O}(\lambda(p)N^{-1/2}p^2)$, while the trace threshold is $\mathcal{O}(\lambda(p)p(n_{\text{Trace}}(n,p))^{-1/2}) = \mathcal{O}(\lambda(p)N^{-1/2}p^2)$ □

For the statistical test $(\mathcal{T}_{ij}, \hat{\Theta}_{ij}, \delta)$ (cf. Eq. (3.14)), if $|\hat{\Theta}_{ij}| \geq t$, then the test rejects the null hypothesis at a significance level δ.

In the simulation study, we set $\mu = 1$, which we have empirically validated to result in a sound test threshold for a wide range of distributions. As discussed below, for a trace threshold on a matrix with condition number $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$, the trace over-estimates Eq. (3.24 A) by at least a factor of $1 + \frac{(p^2 - \binom{p}{2} - 1)\lambda_{\min}}{\lambda_{\max}}$, and the resulting test is therefore valid for distributions for which Eq. (3.24 B) is asymptotically at most $\frac{p^2 - \binom{p}{2} - 1}{\kappa}$ as large as Eq. (3.24 A).

The computation of the statistical test for structure discovery in multivariate graphical models is described in detail in Algorithm 1.

Remark 1. *In the case that ε is larger than the smallest eigenvalue of $\hat{\Theta}$, the test threshold is unbounded and we can never reject the null hypothesis. In this case, additional data are necessary to decrease ε in order to have a non-trivial bound. Theorem 4 guarantees that ε converges to zero as a function of the sample size at a rate $\mathcal{O}(n^{-1/2})$.*

Theorem 8. *For a test with computational cost $\Omega(n^s)$ and a threshold that decreases as $\Omega(n^r)$, our test is asymptotically more powerful in the regime $n \gg p$ whenever $\frac{r}{s} > -\frac{1}{2}$.*

Proof. Our tests have computation $C_{\text{Trace}}(n) \asymp C_{\text{Eig}}(n) \asymp n$. The convergence of our test threshold is $\mathcal{O}(n^{-1/2})$ so for a fixed computational budget N, the test threshold is $\mathcal{O}(N^{-1/2})$. For a test with computational cost $\Omega(n^s)$ and a computational budget N, $\mathcal{O}(N^{1/s})$ samples will be processed. As n^r is decreasing

Algorithm 1 Hypothesis Testing Using a U-statistic estimator for the precision matrix

Require: δ, the significance level of the test; μ, a constant satisfying (3.25); $\mathrm{X} = (X_1, ..., X_p)$ the set of random variables of dimension p with sample size n.

Ensure:

Compute $\hat{\Sigma}$, the unbiased estimator of Σ from X (cf. Def. 2).

Compute $\hat{\Theta} = \hat{\Sigma}^{-1}$, the estimator of the precision matrix.

Compute $U([\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})])$ the upper triangular of the covariance of $U(\hat{\Sigma})$ where (i, j, k, l) vary over the set of p variables (cf. Thm. 3).

Compute

- λ_{max}, the largest eigenvalue of $\mathrm{Cov}(\hat{\Sigma})$, or
- $\mathrm{Tr}[\mathrm{Cov}(\hat{\Sigma})]$, the trace of $\mathrm{Cov}(\hat{\Sigma})$.

Compute one of the two error bounds ε (cf.Eq. (3.30))

- $\varepsilon_{\mathrm{Eig}} = \sqrt{2\lambda_{max}}\Phi^{-1}\left(1 - \delta/2\right)$, or
- $\varepsilon_{\mathrm{Trace}} = \sqrt{2\,\mathrm{Tr}[\mathrm{Cov}(\hat{\Sigma})]}\Phi^{-1}\left(1 - \delta/2\right)$

where Φ is the CDF of a standard normal distribution.

if ε is greater than the smallest eigenvalue of $\hat{\Sigma}$ **then**

  $t = \infty$

else

  Compute the conservative threshold for the two error bound,

  $$t = \mu\sqrt{\sum_{k=1}^{p}\left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2},$$

  where $\hat{\alpha}_k$ is the k-th eigenvalue of the unbiased estimator $\hat{\Sigma}$.

end if

return t.

in n for any consistent test, this implies that the test threshold is $\Omega(N^{r/s})$ which is asymptotically larger than $\mathcal{O}(N^{-1/2})$ whenever $\frac{r}{s} > -\frac{1}{2}$. □

Corollary 2. *Any test that is superlinear must have a threshold that converges faster than $\mathcal{O}(n^{-1/2})$ to be asymptotically more powerful at a fixed computational budget than the tests proposed here.*

4. Simulation Studies

In this section, we demonstrate the soundness and effectiveness of the proposed test which enables one to answer if an edge is significantly present in a graph. This is demonstrated both in terms of experiments on randomly generated Gaussian graphical models with known analytic precision matrices Θ. In all experiments, we have used a significance upper bound of $\delta < 0.05$.

In the simulation, we generated the data X from multivariate Gaussian or Laplace distributions with known analytic precision matrices $\Theta = \Sigma^{-1}$, such that $\Sigma_{ij} = X_i^T X_j / \left(\|X_i\|_2 \|X_j\|_2\right)$ for all $(i, j) \in \{1, ..., p\}$.

1. Multivariate Gaussian distribution

$$f(\mathrm{X}, \Sigma) = \frac{1}{\sqrt{2\pi^p |\Sigma|}} \exp\left\{-\frac{1}{2}\,\mathrm{X}^T\,\Theta\,\mathrm{X}\right\}. \tag{4.1}$$

2. Multivariate Laplace distribution (Gómez et al., 1998)

$$f(\mathrm{X}, \Sigma) = \frac{p\Gamma\left(\frac{p}{2}\right)}{\pi^{\frac{p}{2}}\Gamma\left(1+\frac{p}{\omega}\right)2^{1+\frac{d}{\omega}}|\Sigma|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}\left[\mathrm{X}^T\,\Theta\,\mathrm{X}\right]^{\frac{\omega}{2}}\right\}. \tag{4.2}$$

   For $\omega = 1$, the multivariate Laplace distribution is derived.

In Figure 1, we plot the sample size sample size for 101 regularly spaces values of $n \in [10000, 1010000]$ versus the empirical threshold t_{Eig} and t_{Trace} (cf. Eq. (3.29)) of the test for different numbers of variables p. We clearly distinguish that the threshold t_{Eig} based on the eigenvalue bound in Eq. (3.16) is less than the threshold t_{Trace} based on the trace bound in Eq. (3.18) as predicted by Corollary 1. Furthermore, we see that there is a dependence on the size of the graph, with the bounds growing with the number of variables p.

In Figure 2, we illustrate the inequality of Weyl's Theorem (Thm 5). We show the boxplots of the accurate values of eigenvalues of Θ obtained from the simulation study described. As expected, for a known precision matrix Θ, the eigenvalues $1/\alpha_i, i \in \{1, ...p\}$ is bounded by the two error bounds ε_{Eig} and $\varepsilon_{\text{Trace}}$. As the sample size n increases, the two bounds become tighter.

Then, we compare our edge detection test with the eigenvalue threshold and the trace threshold (*edgeTest-eig* and *edgeTest-tr*) to the Fisher test (*FisherTest*) described in Section 2.2 for different multivariate distributions. The simulations are repeated 100 times to provide statistical significance results.

In Figure 3, we plot the significance level of the test δ against the false positive rate, which refers to the probability of falsely rejecting the null hypothesis for $n = 100000$ and $p = 6$. The diagonal dotted black line indicates that the significance level of different tests is equal to false positive rate. Curves above the diagonal indicate that the test does not obey the semantics of (a bound on) the false positive probability, while a curve under the diagonal indicates that the proposed test is conservative but sound. For the Gaussian distribution (Fig. 3a), the conditional independence test is well calibrated while the proposed test is sound. However, for the Laplace distribution (Fig. 3b), the Fisher test is not valid while the proposed test is sound. Therefore, in Fig. 3c and Fig. 3d, we plot the probability of detecting an edge for all entries on the precision matrix Θ, i.e. when $|\Theta_{ij} - \hat{\Theta}_{ij}| > t$ for all $(i, j) \in \{1, ..., p\}$.

In Figure 4, we compare the power of the tests by plotting the sample size for 101 regularly spaced values of $n \in [10000, 1010000]$ against the power of the test. As expected, in Figs. 4a and 4b, we show that the power of the test increases as the sample size n is increased. In Figs. 4c and 4d, we take into account an effect in the graph in the sense that we want to detect edge only when there is a high correlation between two edges in the graph, i.e. when $|\Theta_i j| > 0.5$ for all $(i, j) \in \{1, ...p\}$.

5. Discussion

We have considered the problem of structure discovery for undirected graphical models in the context of non-Gaussian multivariate distributions, use a concentration bound for U-statistics, leading to two probabilistic bounds t_{Eig} and t_{Trace}. As a baseline, we compare to the Fisher test which is only correct under the assumption of a Gaussian distribution. As shown in the simulation studies, for non-Gaussian distributions, the Fisher test is not calibrated, while alternatively, the proposed test is conservative Among the two probabilistic bounds presented here, the eigenvalue bound is preferred when availability of data is more limited than computation, while t_{Trace} is a competitive test when we have a fixed computational budget N.

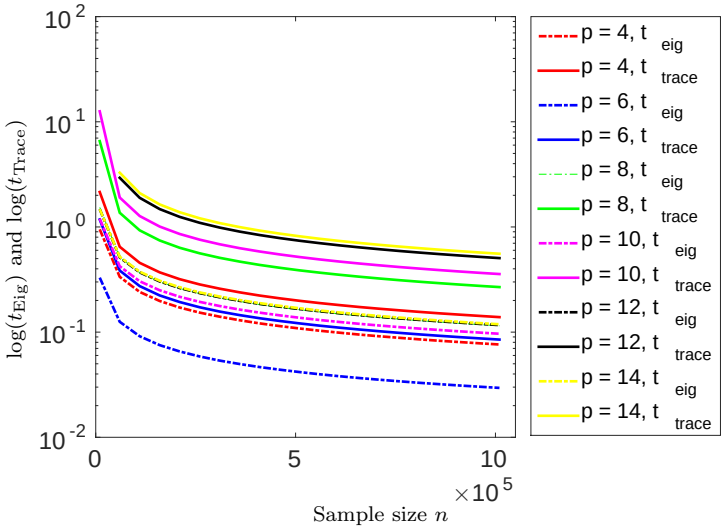
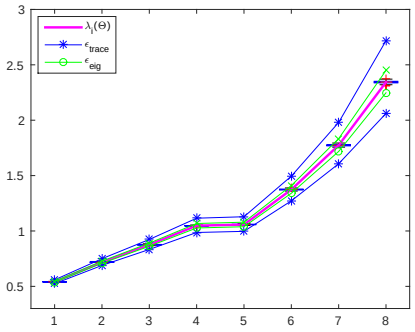


Fig 1: Illustration of the sample size for 101 regularly spaces values of $n \in [10000, 1010000]$ versus the thresholds t_{Eig} and t_{Trace} (Eq. (3.29)). We have plotted both the eigenvalue bound as well as the trace bound (cf. Lemma 1).

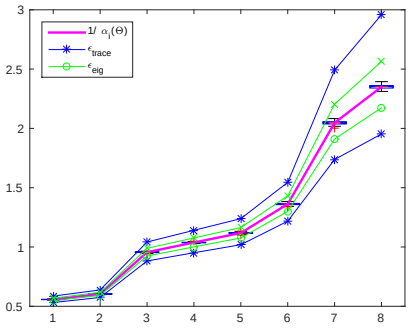
6. Conclusion

In this work, we have constructed a conservative threshold on the absolute value of the precision matrix as a hypothesis test of the presence of an edge in a graphical model. For a wider range of distributions, we have developed a threshold based on a U-statistic empirical estimator of the covariance matrix. This is achieved by probabilistically bounding the distortion of the true covariance matrix, and then using this fixed bound in conjunction with Weyl's theorem to bound the distortion of the precision matrix. These bounds are applicable to the quantification of uncertainty in the magnitude of an effect between variables as measured by the value of the precision matrix, and can also be used to construct a hypothesis test of whether an edge is present in a graphical model by testing for significant deviations from zero. The resulting test asymptotically converges at the same $\frac{1}{\sqrt{n}}$ rate as the U-statistic, which we have additionally verified empirically. We have shown two alternative thresholds, one based on the largest eigenvalue of $\text{Cov}(\hat{\Sigma})$, and a second based on the trace of $\text{Cov}(\hat{\Sigma})$, which strictly upper bounds the first. Given arbitrary computation, we clearly favor the eigenvalue based approach, but for larger graphs with a large number of samples, the tighter threshold yields a test with computational complexity $\mathcal{O}(np^4)$ (due to the requirement of estimating $\mathcal{O}(p^4)$ entries of $\text{Cov}(\hat{\Sigma})$ each of which has linear complexity) while the second has reduced complexity $\mathcal{O}(np^2 + p^3)$ as we need only compute the $\mathcal{O}(p^2)$ diagonal elements of $\text{Cov}(\hat{\Sigma})$. We have shown that



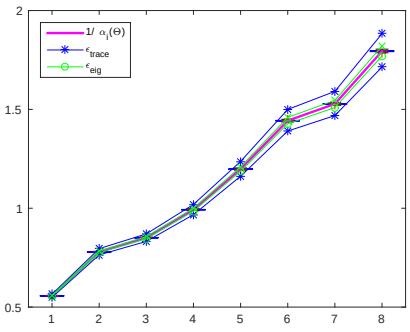
$1/\alpha_i(\Theta)$, for $i \in \{1, ..., 8\}$.

(a) Gaussian dist., $n = 100000$, $p = 6$.



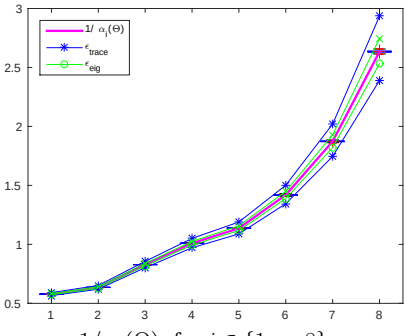
$1/\alpha_i(\Theta)$, for $i \in \{1, ..., 8\}$.

(b) Laplace dist., $n = 100000$, $p = 6$.



$1/\alpha_i(\Theta)$, for $i \in \{1, ..., 8\}$.

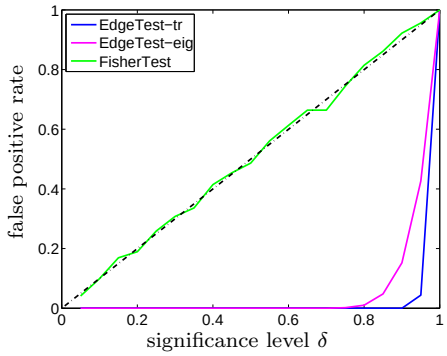
(c) Gaussian dist, $n = 500000$, $p = 6$.



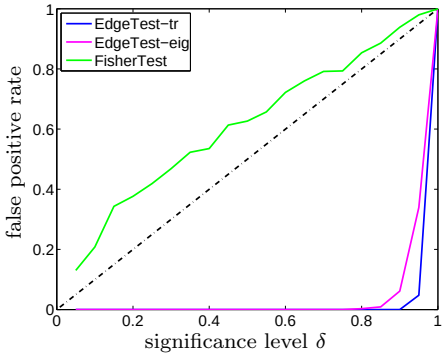
$1/\alpha_i(\Theta)$, for $i \in \{1, ..., 8\}$.

(d) Laplace dist., $n = 500000$, $p = 6$.

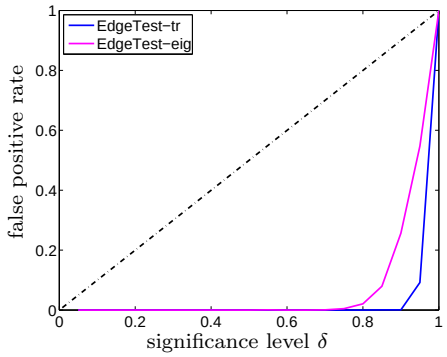
Fig 2: For a known analytic precision matrix Θ of size $p = 8$ and for two different sample sizes, we show the boxplots of accuracy values of eigenvalues of 200 estimates matrices $\hat{\Theta}$ for the Gaussian (Figs 2a, 2c) and Laplace (Figs 2b, 2d) distributions with normalized data. In pink, we plot the true eigenvalue of Θ and in green and blue, we plot the upper and lower bound given by Weyl's theorem. As n grows, we see that the bound more closely constrains the true eigenvalues of Θ.



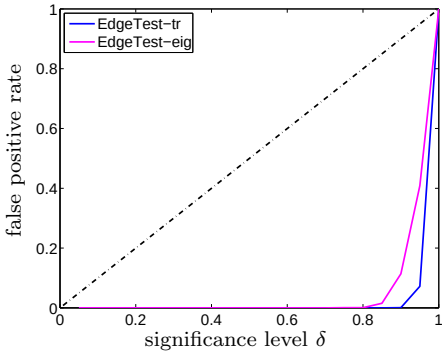
(a) Gaussian dist., $n = 100000$, $p = 6$.



(b) Laplace dist., $n = 100000$, $p = 6$.

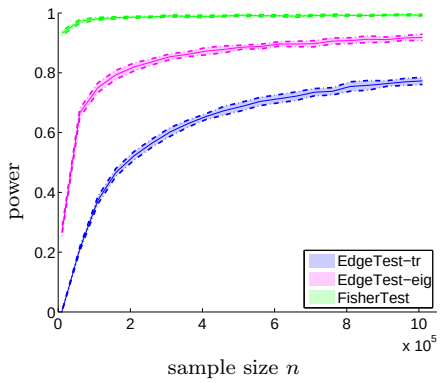


(c) Gaussian dist., $n = 100000$, $p = 6$.

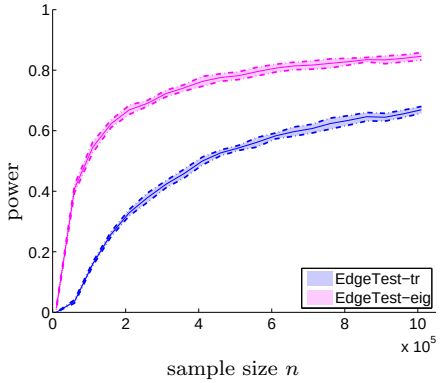


(d) Laplace dist., $n = 100000$, $p = 6$.

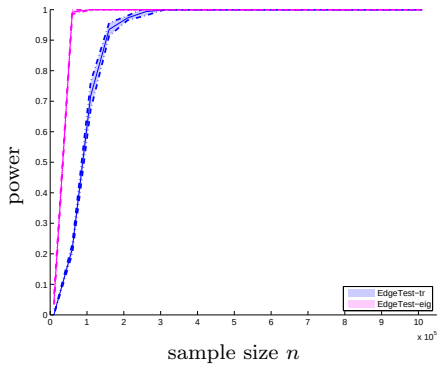
Fig 3: We compare the false positive rate for the proposed test and the Fisher test. For the Gaussian distribution (Fig 3a), the curves show that the Fisher test is well calibrated and that the proposed test is conservative (below the diagonal). Furthermore, for the Laplace distribution (Fig 3b), the Fisher test does not obey the semantics of a bound on δ (the curve is above the diagonal) while by contrast, the proposed test remains conservative and sound. In Fig 3c and Fig. 3d, we compare the rate of violating a bound on the true precision matrix as a function of δ, i.e when $|\hat{\Theta}_{ij} - \Theta_{ij}| > t$ for an (i, j) in $U(\Theta)$.



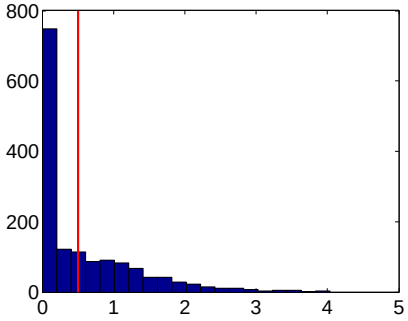
(a) Gaussian dist.



(b) Laplace dist.



(c) Laplace dist.



(d) Histogram of $|\Theta_{ij}|$ for 100 random graphs drawn from a Laplace distribution.

Fig 4: As a function of the sample size n, we compare the power of the proposed test and the Fisher test for the Gaussian distribution (Fig. 4a) and for the Laplace distribution (Fig. 4b). In Fig. 4c, we plot the power of the proposed test when we reject the null hypothesis and when $|\Theta_{ij}| > 0.5$ (see histogram 4d). The shaded region indicates the standard error estimated from multiple repetitions. The proposed tests are more generally applicable than the Fisher test, and have high power for edges with strong effects, i.e. those which are most important to detect and model.

this reduced complexity makes the trace bound competitive when computation rather than data availability is the restrictive factor.

The construction of the test threshold has upper bounded the $\|\cdot\|_2$ matrix norm with the Frobenius norm, which leads to a systematic overestimation of the threshold proportional to the size of the graph. This is clearly demonstrated in the simulation study section. We have taken the approach of probabilistically bounding the distortion of the covariance, and then, given this fixed bound, uniformly bounding the distortion of the precision matrix. It may be of interest to consider a non-uniform bound to reduce the growth of the bound in the number of variables.

Simulation studies show that the test successfully recovers the structure of undirected graphical models given a sufficient number of samples. The sample complexity increases with the size of the smallest non-zero entry of Θ as well as with the number of variables in the model. Figure 1 demonstrates that the bound tends to grow with the size of the graph for a fixed sample size, while the size of the non-zero entries follows the same distribution in these experiments. Nevertheless, large values of Θ can be recovered with significance even in these cases. The fact that the test was able to compute correct results even for $n = 10^6$ and $p = 14$ in a short time demonstrates the scalability and soundness of the approach.

Appendix A: Derivation of the covariance of the *U*-statistics for the covariance matrix

In this appendix, we show the details of the derivation of Theorem 2. We derive low variance, unbiased estimates of the covariance between two U-statistics estimates $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$, where (i, j, k, l) range over each of the d variates in a covariance matrix $\hat{\Sigma}$. We note h and g the corresponding kernel of order 2 for $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$, where

$$h(u_1, u_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{j_1} - X_{j_2}\right), \text{with } u_r = (X_{i_r}, X_{j_r})^T \tag{A.1}$$

$$g(v_1, v_2) = \frac{1}{2}\left(X_{k_1} - X_{k_2}\right)\left(X_{l_1} - X_{l_2}\right), \text{with } v_r = (X_{k_r}, X_{l_r})^T. \tag{A.2}$$

Then, the covariance $\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$ for the two U-statistics $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$ is

$$\begin{aligned}\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) &= \binom{m}{2}^{-1}\left(2(m-2)\zeta_1 + \zeta_2\right) \\ &= \binom{m}{2}^{-1}\left(2(m-2)\zeta_1\right) + \mathcal{O}(m^{-2})\end{aligned} \tag{A.3}$$

where $\zeta_1 = \mathrm{Cov}\left(\mathrm{E}_{u_2}[h(u_1, u_2)], \mathrm{E}_{v_2}[g(v_1, v_2)]\right)$.

Depending on the equality and inequality of these four index variables, the empirical covariance estimate takes a different kernel form. We have employed a computer assisted proof to determine that there are seven different forms and that each of the unique $\binom{p^2-\binom{p}{2}}{2}$ entries in $\mathrm{Cov}(\hat{\Sigma})$ (cf. Eq. (3.4)) can be mapped to one of these seven cases by a simple variable substitution.

In the sequel, we first describe the algorithm that determines the seven cases (Sec. A.1), we derive empirical estimators for each of these seven cases (Sec. A.2) and show that in all cases we have linear computation time in the number of samples (Sec. A.3).

A.1. Description of the algorithm providing the seven cases

We formally described the algorithm that provided us 7 cases for the derivation of $\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$ of Theorem 2, where (i, j, k, l) vary over the set of d variables.

Enumeration First, we enumerate all configurations of $\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$, which can be encoded as a non-unique assignment matrix of variables i, j, k, l to instantiated variables (a, b, c, d). For a fixed assignment of i to variable a, we can list all possible assignments of the 3 remaining variables (j, k, l) to any (a, b, c, d). Naïvely, we have 4^3 possible assignments, but many of them will be equivalent by variable substitution. To test whether two forms are equivalent, it is sufficient to test a reduced form for equality.

Reduced Form We map a variable assignment to a reduced form by re-labeling variables sorted by the number of occurrences, which reduces the number of possible matches up-to non-uniqueness of the mapping due to equal numbers of variable occurrences. This ambiguity is then resolved by testing for symmetries.

Symmetry Symmetry of the covariance operator brings the following equally that we take into consideration in testing for equivalence:

$$\begin{aligned}\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) = \mathrm{Cov}(\hat{\Sigma}_{kl}, \hat{\Sigma}_{ij}) = \mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{lk}) = \mathrm{Cov}(\hat{\Sigma}_{lk}, \hat{\Sigma}_{ij}) \\ = \mathrm{Cov}(\hat{\Sigma}_{lk}, \hat{\Sigma}_{ji}) = \mathrm{Cov}(\hat{\Sigma}_{ji}, \hat{\Sigma}_{kl}) = \mathrm{Cov}(\hat{\Sigma}_{ji}, \hat{\Sigma}_{lk}).\end{aligned} \tag{A.4}$$

The algorithm outputs each variable assignment that is not equivalent by variable substitution to any previously enumerated assignment. Open source code for the computer assisted proof is available at https://github.com/wbounliphone/Ustatistics_Approach_For_SD.

The seven different cases are enumerated in Table 2.

Cases	Indices	Correspondence
1	$i \neq j, k, l;\ j \neq k, l;\ k \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$
2	$i = j;\ j \neq k, l;\ k = l$	$\mathrm{Cov}(\hat{\Sigma}_{ii}, \hat{\Sigma}_{kk})$
3	$i = j;\ j \neq k, l;\ k \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ii}, \hat{\Sigma}_{kl})$
4	$i = k;\ j \neq i, k, l;\ k \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{il})$
5	$i = k;\ i \neq j;\ j = l;$	$\mathrm{Var}(\hat{\Sigma}_{ij})$
6	$i = j = k;\ i \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ii}, \hat{\Sigma}_{il})$
7	$i = j, k, l$	$\mathrm{Var}(\hat{\Sigma}_{ii})$

TABLE 2
Enumeration and correspondence of the seven cases.

A.2. The seven exhaustive cases

We now derive linear-time finite-sample estimates of the covariance for each of the seven cases.

Notation

- $\overline{XYUV} = \mathrm{E}[XYUV]$
- $\overline{XYZ} = \mathrm{E}[XYZ]$
- $\overline{XY} = \mathrm{E}[XY]$
- $\overline{X} = \mathrm{E}[X]$
- $\overline{XYUV}\ \overline{X} = \mathrm{E}[XYUV] \times \mathrm{E}[X]$

A.2.1. Case 1: $i \neq j, k, l;\ j \neq k, l;\ k \neq l$

The kernels are

$$h(u_1, u_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{j_1} - X_{j_2}\right); \qquad g(v_1, v_2) = \frac{1}{2}\left(X_{k_1} - X_{k_2}\right)\left(X_{l_1} - X_{l_2}\right)$$

$$\mathrm{E}_{u_2}[h(u_1, u_2)] = \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{j_1} - \overline{X_j}\right); \mathrm{E}_{u_2}[g(v_1, v_2)] = \frac{1}{2}\left(X_{k_1} - \overline{X_k}\right)\left(X_{l_1} - \overline{X_l}\right)$$

$$
\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(X_{i_1}-\overline{X_i}\right)\left(X_{j_1}-\overline{X_j}\right),\frac{1}{2}\left(X_{k_1}-\overline{X_k}\right)\left(X_{l_1}-\overline{X_l}\right)\right] \qquad \text{(A.5)}\\
&= \frac{1}{4}\Bigg\{\mathrm{Cov}\left[X_{i_1}X_{j_1}-\overline{X_i}X_{j_1}-X_{i_1}\overline{X_j};X_{k_1}X_{l_1}-\overline{X_k}X_{l_1}-X_{k_1}\overline{X_l}\right]\Bigg\}\\
&= \frac{1}{4}\Bigg\{\mathrm{E}_{u_1}\left[X_{i_1}X_{j_1}X_{k_1}X_{l_1}-\overline{X_i}X_{j_1}X_{k_1}X_{l_1}-X_{i_1}\overline{X_j}X_{k_1}X_{l_1}\right.\\
&\qquad\qquad -X_{i_1}X_{j_1}\overline{X_k}X_{l_1}+\overline{X_i}X_{j_1}\overline{X_k}X_{l_1}+X_{i_1}\overline{X_j}\ \overline{X_k}X_{l_1}\\
&\qquad\qquad \left.-X_{i_1}X_{j_1}X_{k_1}\overline{X_l}+\overline{X_i}X_{j_1}X_{k_1}\overline{X_l}+X_{i_1}\overline{X_j}X_{k_1}\overline{X_l}\right]\\
&\qquad -\mathrm{E}_{u_1}\left[X_{i_1}X_{j_1}-\overline{X_i}X_{j_1}-X_{i_1}\overline{X_j}\right]\mathrm{E}_{u_1}\left[X_{k_1}X_{l_1}-\overline{X_k}X_{l_1}-X_{k_1}\overline{X_l}\right]\Bigg\}\\
&= \frac{1}{4}\Bigg\{\overline{X_iX_jX_kX_l}-\overline{X_i}\ \overline{X_jX_kX_l}-\overline{X_j}\ \overline{X_iX_kX_l}\\
&\qquad -\overline{X_k}\ \overline{X_iX_jX_l}+\overline{X_i}\ \overline{X_k}\ \overline{X_jX_l}+\overline{X_j}\ \overline{X_k}\ \overline{X_iX_l}\\
&\qquad -\overline{X_iX_jX_k}\ \overline{X_l}+\overline{X_i}\ \overline{X_l}\ \overline{X_jX_k}+\overline{X_j}\ \overline{X_l}\ \overline{X_iX_k}\\
&\qquad -\left(\overline{X_iX_j}-2\ \overline{X_i}\ \overline{X_j}\right)\left(\overline{X_kX_l}-2\ \overline{X_k}\ \overline{X_l}\right)\Bigg\}
\end{aligned}
$$

A.2.2. Case 2: $i=j;\ j\neq k,l;\ k=l$

The kernels are

$$
\begin{aligned}
h(u_1,u_2) &= \frac{1}{2}\left(X_{i_1}-X_{i_2}\right)^2; & g(v_1,v_2) &= \frac{1}{2}\left(X_{k_1}-X_{k_2}\right)^2\\
\mathrm{E}_{u_2}[h(u_1,u_2)] &= \frac{1}{2}\left(X_{i_1}-\overline{X_i}\right)^2; & \mathrm{E}_{u_2}[g(v_1,v_2)] &= \frac{1}{2}\left(X_{k_1}-\overline{X_k}\right)^2
\end{aligned}
$$

Then, we have

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(X_{i_1}-\overline{X_i}\right)^2;\frac{1}{2}\left(X_{k_1}-\overline{X_k}\right)^2\right] \qquad \text{(A.6)}\\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[X_{i_1}^2-2X_{i_1}\overline{X_i};X_{k_1}^2-2X_{k_1}\overline{X_k}\right]\Big\}\\
&= \frac{1}{4}\Big\{\mathrm{E}_{u_1}\left[X_{i_1}^2X_{k_1}^2-2X_{i_1}\overline{X_i}X_{k_1}^2-2X_{i_1}^2X_{k_1}\overline{X_k}+4X_{i_1}\overline{X_i}X_{k_1}\overline{X_k}\right]\\
&\quad -\mathrm{E}_{u_1}\left[X_{i_1}^2-2X_{i_1}\overline{X_i}\right]\mathrm{E}_{u_1}\left[X_{k_1}^2-2X_{k_1}\overline{X_k}\right]\Big\}\\
&= \frac{1}{4}\Big\{\overline{X_i^2X_k^2}-2\ \overline{X_i}\ \overline{X_iX_k^2}-2\ \overline{X_i^2X_{k_1}}\ \overline{X_k}+4\overline{X_iX_k}\ \overline{X_i}\ \overline{X_k}\\
&\quad -\left(\overline{X_i^2}-2\ \overline{X_i}^2\right)\left(\overline{X_k^2}-2\ \overline{X_k}^2\right)\Big\}
\end{aligned}$$

A.2.3. Case 3: $i=j;\ j\neq k,l;\ k\neq l$

The kernels are

$$\begin{aligned}
h(u_1,u_2) &= \frac{1}{2}\left(X_{i_1}-X_{i_2}\right)^2; \qquad & g(v_1,v_2) &= \frac{1}{2}\left(X_{k_1}-X_{k_2}\right)\left(X_{l_1}-X_{l_2}\right)\\
\mathrm{E}_{u_2}[h(u_1,u_2)] &= \frac{1}{2}\left(X_{i_1}-c\right)^2; \qquad & \mathrm{E}_{u_2}[g(v_1,v_2)] &= \frac{1}{2}\left(X_{k_1}-\overline{X_k}\right)\left(X_{l_1}-\overline{X_l}\right)
\end{aligned}$$

Then, we have

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(X_{i_1}-\overline{X_i}\right)^2;\frac{1}{2}\left(X_{k_1}-\overline{X_k}\right)\left(X_{l_1}-\overline{X_l}\right)\right] \qquad \text{(A.7)}\\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[X_{i_1}^2-2X_{i_1}\overline{X_i};X_{k_1}X_{l_1}-\overline{X_k}X_{l_1}-X_{k_1}\overline{X_l}\right]\Big\}\\
&= \frac{1}{4}\Big\{\mathrm{E}_{u_1}\left[X_{i_1}^2X_{k_1}X_{l_1}-2X_{i_1}\overline{X_i}X_{k_1}X_{l_1}-X_{i_1}^2\overline{X_k}X_{l_1}\right.\\
&\qquad \left.+2X_{i_1}\overline{X_i}\ \overline{X_k}X_{l_1}-X_{i_1}^2X_{k_1}\overline{X_l}+2X_{i_1}\overline{X_i}X_{k_1}\overline{X_l}\right]\\
&\quad -\mathrm{E}_{u_1}\left[X_{i_1}^2-2X_{i_1}\overline{X_i}\right]\mathrm{E}_{u_1}\left[X_{k_1}X_{l_1}-\overline{X_k}X_{l_1}-X_{k_1}\overline{X_l}\right]\Big\}\\
&= \frac{1}{4}\Big\{\overline{X_i^2X_kX_l}-2\ \overline{X_iX_kX_l}\ \overline{X_i}-\overline{X_i^2X_l}\ \overline{X_k}\\
&\qquad +2\ \overline{X_iX_l}\ \overline{X_i}\ \overline{X_k}-\overline{X_i^2X_{k_1}}\ \overline{X_l}+2\ \overline{X_iX_k}\ \overline{X_i}\ \overline{X_l}\\
&\quad -\left(\overline{X_i^2}-2\ \overline{X_i}^2\right)\left(\overline{X_kX_l}-2\ \overline{X_k}\ \overline{X_l}\right)\Big\}
\end{aligned}$$

A.2.4. Case 4: $i = k$; $j \neq i, k, l$; $k \neq l$

The kernels are

$$h(u_1, u_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{j_1} - X_{j_2}\right); \qquad g(v_1, v_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{l_1} - X_{l_2}\right)$$
$$\mathrm{E}_{u_2}[h(u_1, u_2)] = \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{j_1} - \overline{X_j}\right); \mathrm{E}_{u_2}[g(v_1, v_2)] = \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{l_1} - \overline{X_l}\right)$$

Then, we have

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{j_1} - \overline{X_j}\right); \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{l_1} - \overline{X_l}\right)\right] \qquad \text{(A.8)}\\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[X_{i_1}X_{j_1} - \overline{X_i}X_{j_1} - X_{i_1}\overline{X_j}; X_{i_1}X_{l_1} - \overline{X_i}X_{l_1} - X_{i_1}\overline{X_l}\right]\Big\}\\
&= \frac{1}{4}\Big\{\mathrm{E}_{u_1}\left[X_{i_1}^2 X_{j_1}X_{l_1} - \overline{X_i}X_{j_1}X_{i_1}X_{l_1} - X_{i_1}^2\overline{X_j}X_{l_1}\right.\\
&\qquad - X_{i_1}X_{j_1}\overline{X_i}X_{l_1} + \overline{X_i}^2\ X_{j_1}X_{l_1} + X_{i_1}\overline{X_j}\ \overline{X_i}X_{l_1}\\
&\qquad \left. - X_{i_1}^2X_{j_1}\overline{X_l} + \overline{X_i}X_{j_1}X_{i_1}\overline{X_l} + X_{i_1}^2\overline{X_jX_l}\right]\\
&\quad - \mathrm{E}_{u_1}\left[X_{i_1}X_{j_1} - \overline{X_i}X_{j_1} - X_{i_1}\overline{X_j}\right]\mathrm{E}_{u_1}\left[X_{i_1}X_{l_1} - \overline{X_i}X_{l_1} - X_{i_1}\overline{X_l}\right]\Big\}\\
&= \frac{1}{4}\Big\{\overline{X_{i_1}^2X_{j_1}X_{l_1}} - \overline{X_i}\ \overline{X_{j_1}X_{i_1}X_{l_1}} - \overline{X_{i_1}^2X_{l_1}}\ \overline{X_j}\\
&\qquad - \overline{X_{i_1}X_{j_1}X_{l_1}}\ \overline{X_i} + \overline{X_i}^2\ \overline{X_{j_1}X_{l_1}} + \overline{X_{i_1}X_{l_1}}\ \overline{X_j}\ \overline{X_i}\\
&\qquad - \overline{X_{i_1}^2X_{j_1}}\ \overline{X_l} + \overline{X_i}\ \overline{X_{j_1}X_{i_1}}\ \overline{X_l} + \overline{X_{i_1}^2}\ \overline{X_j}\ \overline{X_l}]\\
&\quad - \left(\overline{X_iX_j} - 2\ \overline{X_i}\ \overline{X_j}\right)\left(\overline{X_iX_l} - 2\ \overline{X_i}\ \overline{X_l}\right)\Big\}
\end{aligned}$$

A.2.5. Case 5: $i = k$; $i \neq j$; $j = l$;

$$h(u_1, u_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{j_1} - X_{j_2}\right); \qquad g(v_1, v_2) = h(u_1, u_2)$$
$$\mathrm{E}_{u_2}[h(u_1, u_2)] = \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{j_1} - \overline{X_j}\right); \qquad \mathrm{E}_{u_2}[g(v_1, v_2)] = \mathrm{E}_{u_2}[h(u_1, u_2)]$$

Then, we have

$$\begin{aligned}
\zeta_1 &= \mathrm{Var}\left[\frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{j_1} - \overline{X_j}\right)\right] \qquad \text{(A.9)}\\
&= \frac{1}{4}\Bigg\{\mathrm{Var}\left[X_{i_1}X_{j_1} - \overline{X_i}X_{j_1} - X_{i_1}\overline{X_j}\right]\Bigg\}\\
&= \frac{1}{4}\Bigg\{\mathrm{E}_{u_1}\left[(X_{i_1}X_{j_1} - \overline{X_i}X_{j_1} - X_{i_1}\overline{X_j})^2\right] - \mathrm{E}_{u_1}\left[X_{i_1}X_{j_1} - \overline{X_i}X_{j_1} - X_{i_1}\overline{X_j}\right]^2\Bigg\}\\
&= \frac{1}{4}\Bigg\{\mathrm{E}_{u_1}\left[X_{i_1}^2X_{j_1}^2 - 2X_{i_1}X_{j_1}^2\overline{X_i} + \overline{X_i}^2X_{j_1}^2 - 2X_{i_1}^2X_{j_1}\overline{X_j} + 2\overline{X_i}X_{j_1}X_{i_1}\overline{X_j} + X_{i_1}^2\overline{X_j}^2\right]\\
&\qquad - \left(\overline{X_iX_j} - 2(\overline{X_i}\ \overline{X_j})\right)^2\Bigg\}\\
&= \frac{1}{4}\Bigg\{\overline{X_i^2X_j^2} - 2\overline{X_iX_j^2}\ \overline{X_i} + \overline{X_i}^2\ \overline{X_j^2} - 2\overline{X_i^2X_j}\ \overline{X_j} + 2\overline{X_i}\ \overline{X_j}\ \overline{X_jX_i} + \overline{X_i^2}\ \overline{X_j}^2\\
&\qquad - \left(\overline{X_iX_j} - 2(\overline{X_i}\ \overline{X_j})\right)^2\Bigg\}
\end{aligned}$$

A.2.6. Case 6: $i = j = k$; $i \neq l$

The kernels are

$$\begin{aligned}
h(u_1, u_2) &= \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)^2; \qquad & g(v_1, v_2) &= \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)\left(X_{l_1} - X_{l_2}\right)\\
\mathrm{E}_{u_2}[h(u_1, u_2)] &= \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)^2; \qquad & \mathrm{E}_{u_2}[g(v_1, v_2)] &= \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{l_1} - \overline{X_l}\right)
\end{aligned}$$

Then, we have

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)^2; \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)\left(X_{l_1} - \overline{X_l}\right)\right] \qquad \text{(A.10)}\\
&= \frac{1}{4}\Bigg\{\mathrm{Cov}\left[X_{i_1}^2 - 2X_{i_1}\overline{X_i}; X_{i_1}X_{l_1} - \overline{X_i}X_{l_1} - X_{i_1}\overline{X_l}\right]\Bigg\}\\
&= \frac{1}{4}\Bigg\{\mathrm{E}_{u_1}\left[X_{i_1}^2X_{i_1}X_{l_1} - 2X_{i_1}\overline{X_i}X_{i_1}X_{l_1} - X_{i_1}^2\overline{X_i}X_{l_1}\right.\\
&\qquad \left. + 2X_{i_1}\overline{X_i}\ \overline{X_i}X_{l_1} - X_{i_1}^2X_{i_1}\overline{X_l} + 2X_{i_1}\overline{X_i}X_{i_1}\overline{X_l}\right]\\
&\qquad - \mathrm{E}_{u_1}\left[X_{i_1}^2 - 2X_{i_1}\overline{X_i}\right]\mathrm{E}_{u_1}\left[X_{i_1}X_{l_1} - \overline{X_i}X_{l_1} - X_{i_1}\overline{X_l}\right]\Bigg\}\\
&= \frac{1}{4}\Bigg\{\overline{X_i^3X_l} - 3\ \overline{X_i^2X_l}\ \overline{X_i} + 2\ \overline{X_iX_l}\ \overline{X_i}^2 - \overline{X_i^3}\ \overline{X_l} + 2\ \overline{X_i^2}\ \overline{X_i}\ \overline{X_l}\\
&\qquad - \left(\overline{X_i^2} - 2\ \overline{X_i}^2\right)\left(\overline{X_iX_l} - 2\ \overline{X_i}\ \overline{X_l}\right)\Bigg\}
\end{aligned}$$

A.2.7. Case 7: $i = j, k, l$

The kernels are

$$h(u_1, u_2) = \frac{1}{2}\left(X_{i_1} - X_{i_2}\right)^2; \qquad g(v_1, v_2) = h(u_1, u_2)$$
$$\mathrm{E}_{u_2}[h(u_1, u_2)] = \frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)^2; \qquad \mathrm{E}_{u_2}[g(v_1, v_2)] = \mathrm{E}_{u_2}[h(u_1, u_2)]$$

Then, we have

$$\begin{aligned}
\zeta_1 &= \mathrm{Var}\left[\frac{1}{2}\left(X_{i_1} - \overline{X_i}\right)^2\right] \\
&= \frac{1}{4}\mathrm{Var}\left[X_{i_1}^2 - 2X_{i_1}\overline{X_i}\right] \\
&= \frac{1}{4}\left\{\mathrm{E}_{u_1}\left[\left(X_{i_1}^2 - 2X_{i_1}\overline{X_i}\right)^2\right] - \mathrm{E}_{u_1}\left[X_{i_1}^2 - 2X_{i_1}\overline{X_i}\right]^2\right\} \\
&= \frac{1}{4}\left\{\overline{X_i^4} - 4\overline{X_i^3}\ \overline{X_i} + 4\overline{X_i^2}\ \overline{X_i}^2 - \left(\overline{X_i^2} - 2\overline{X_i}^2\right)^2\right\}
\end{aligned} \tag{A.11}$$

A.3. Derivation in $\mathcal{O}(n)$ time for all terms

In section A.2, all terms are in the form of E[X],E[XY],E[XYZ] and E[$XYUV$] and can be computed in $\mathcal{O}(n)$ as following

$$\mathrm{E}[X] = \frac{1}{m}\sum_{q=1}^{n} X_q \tag{A.12}$$

$$\mathrm{E}[XY] = \frac{1}{m}\sum_{q=1}^{n} X_q \odot Y_q \tag{A.13}$$

$$\mathrm{E}[XYZ] = \frac{1}{m}\sum_{q=1}^{n} X_q \odot Y_q \odot Z_q \tag{A.14}$$

$$\mathrm{E}[XYUV] = \frac{1}{m}\sum_{q=1}^{n} X_q \odot Y_q \odot U_q \odot V_q \tag{A.15}$$

References

Bach, F. R. and M. I. Jordan (2003): "Kernel independent component analysis," *The Journal of Machine Learning Research*, 3, 1–48.

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

Dawid, A. P. (1979): "Conditional independence in statistical theory," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–31.

Dempster, A. P. (1972): "Covariance selection," *Biometrics*, 157–175.

Fisher, R. A. (1924): "The distribution of the partial correlation coefficient," *Metron*, 3, 329–332.

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

Fukumizu, K., F. R. Bach, and M. I. Jordan (2009): "Kernel dimension reduction in regression," *The Annals of Statistics*, 37, 1871–1905.

Fukumizu, K., A. Gretton, X. Sun, and B. Schölkopf (2007): "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems*, vol. 20, 489–496.

Gómez, E., M. Gomez-Viilegas, and J. Marin (1998): "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics-Theory and Methods*, 27, 589–600.

Gretton, A., O. Bousquet, A. J. Smola, and B. Schölkopf (2005): "Measuring statistical dependence with Hilbert-Schmidt norms," in *Algorithmic Learning Theory (ALT)*, 63–77.

Gretton, A., R. Herbrich, and A. J. Smola (2003): "The kernel mutual information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 4, IV–880.

G'Sell, M. G., J. Taylor, and R. Tibshirani (2013): "Adaptive testing for the graphical lasso," *arXiv:1307.4765*.

Heller, R., Y. Heller, and M. Gorfine (2012): "A consistent multivariate test of association based on ranks of distances," *Biometrika*, ass070.

Hoeffding, W. (1948): "A class of statistics with asymptotically normal distribution," *The Annals of Mathematical Statistics*, 293–325.

Janková, J. and S. van de Geer (2015): "Confidence intervals for high-dimensional inverse covariance estimation," *Electronic Journal of Statistics*, 9, 1205–1229.

Jensen, F. V. (1996): *An introduction to Bayesian networks*, UCL Press.

Kendall, M. G. (1938): "A new measure of rank correlation," *Biometrika*, 81–93.

——— (1946): *The advanced theory of statistics*, C. Griffin.

Koller, D. and N. Friedman (2009): *Probabilistic graphical models: Principles and techniques*, MIT Press.

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

Lee, A. J. (1990): *U-statistics: Theory and practice*, CRC Press.

Lehmann, E. L. (1999): *Elements of large-sample theory*, Springer.

Li, H. and J. Gui (2006): "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, 7, 302–317.

Lockhart, R., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2014):

"A significance test for the lasso," *Annals of statistics*, 42, 413.

Loh, P.-L. and M. J. Wainwright (2013): "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses," *The Annals of Statistics*, 41, 3022–3049.

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

Neapolitan, R. E. (2004): *Learning Bayesian networks*, Prentice Hall.

Pearl, J. (2014): *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann.

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

Rényi, A. (1961): "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 547–561.

Roverato, A. and J. Whittaker (1996): "Standard errors for the parameters of graphical Gaussian models," *Statistics and Computing*, 6, 297–302.

Schäfer, J. and K. Strimmer (2005): "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, 4.

Serfling, R. J. (2009): *Approximation theorems of mathematical statistics*, John Wiley & Sons.

Spearman, C. (1904): "The proof and measurement of association between two things," *The American journal of psychology*, 15, 72–101.

Speed, T. P. and H. Kiiveri (1986): "Gaussian Markov distributions over finite graphs," *The Annals of Statistics*, 138–150.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007): "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, 35, 2769–2794.

Temlyakov, V. (2011): *Greedy approximation*, Cambridge University Press.

Tsallis, C. (1988): "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, 52, 479–487.

Weyl, H. (1912): "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)," *Mathematische Annalen*, 71, 441–479.

Whittaker, J. (2009): *Graphical models in applied multivariate statistics*, Wiley Publishing.

Xia, N., Y. Qin, and Z. Bai (2013): "Convergence rates of eigenvector empirical spectral distribution of large dimensional sample covariance matrix," *The Annals of Statistics*, 41, 2572–2607.

Zhang, K., J. Peters, D. Janzing, and B. Schölkopf (2011): "Kernel-based Conditional Independence Test and Application in Causal Discovery," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 804–813.