High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence

Pradeep Ravikumar[†]
pradeepr@stat.berkeley.edu

Martin J. Wainwright[†,♯]
wainwrig@stat.berkeley.edu

Garvesh Raskutti[†]
garveshr@stat.berkeley.edu

Bin Yu[†,♯]
binyu@stat.berkeley.edu

Department of Statistics[†], and
Department of Electrical Engineering and Computer Sciences[♯]
University of California, Berkeley
Berkeley, CA 94720

November 21, 2008

Abstract

Given i.i.d. observations of a random vector $X \in \mathbb{R}^p$, we study the problem of estimating both its covariance matrix Σ^*, and its inverse covariance or concentration matrix $\Theta^* = (\Sigma^*)^{-1}$. We estimate Θ^* by minimizing an ℓ_1-penalized log-determinant Bregman divergence; in the multivariate Gaussian case, this approach corresponds to ℓ_1-penalized maximum likelihood, and the structure of Θ^* is specified by the graph of an associated Gaussian Markov random field. We analyze the performance of this estimator under high-dimensional scaling, in which the number of nodes in the graph p, the number of edges s and the maximum node degree d, are allowed to grow as a function of the sample size n. In addition to the parameters (p, s, d), our analysis identifies other key quantities that control rates: (a) the ℓ_∞-operator norm of the true covariance matrix Σ^*; and (b) the ℓ_∞ operator norm of the sub-matrix Γ^*_{SS}, where S indexes the graph edges, and $\Gamma^* = (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$; and (c) a mutual incoherence or irrepresentability measure on the matrix Γ^* and (d) the rate of decay $1/f(n, \delta)$ on the probabilities $\{|\widehat{\Sigma}^n_{ij} - \Sigma^*_{ij}| > \delta\}$, where $\widehat{\Sigma}^n$ is the sample covariance based on n samples. Our first result establishes consistency of our estimate $\widehat{\Theta}$ in the elementwise maximum-norm. This in turn allows us to derive convergence rates in Frobenius and spectral norms, with improvements upon existing results for graphs with maximum node degrees $d = o(\sqrt{s})$. In our second result, we show that with probability converging to one, the estimate $\widehat{\Theta}$ correctly specifies the zero pattern of the concentration matrix Θ^*. We illustrate our theoretical results via simulations for various graphs and problem parameters, showing good correspondences between the theoretical predictions and behavior in simulations.

1 Introduction

The area of high-dimensional statistics deals with estimation in the "large p, small n" setting, where p and n correspond, respectively, to the dimensionality of the data and the sample size. Such high-dimensional problems arise in a variety of applications, among them remote sensing, computational biology and natural language processing, where the model dimension may be comparable or substantially larger than the sample size. It is well-known that such high-dimensional scaling can lead to dramatic breakdowns in many classical procedures. In the absence of additional model assumptions, it is frequently impossible to obtain consistent procedures when $p \gg n$. Accordingly, an active line of statistical research is based on imposing various restrictions on the model—-for instance, sparsity, manifold structure, or graphical model structure—-and then studying the scaling behavior of different estimators as a function of sample size n, ambient dimension p and additional parameters related to these structural assumptions.

In this paper, we study the following problem: given n i.i.d. observations $\{X^{(k)}\}_{k=1}^n$ of a zero mean random vector $X \in \mathbb{R}^p$, estimate both its covariance matrix Σ^*, and its inverse covariance or concentration matrix $\Theta^* := \left(\Sigma^*\right)^{-1}$. Perhaps the most natural candidate for estimating Σ^* is the empirical sample covariance matrix, but this is known to behave poorly in high-dimensional settings. For instance, when $p/n \to c > 0$, and the samples are drawn i.i.d. from a multivariate Gaussian distribution, neither the eigenvalues nor the eigenvectors of the sample covariance matrix are consistent estimators of the population versions [14, 15]. Accordingly, many regularized estimators have been

proposed to estimate the covariance or concentration matrix under various model assumptions. One natural model assumption is that reflected in shrinkage estimators, such as in the work of Ledoit and Wolf [16], who proposed to shrink the sample covariance to the identity matrix. An alternative model assumption, relevant in particular for time series data, is that the covariance or concentration matrix is banded, meaning that the entries decay based on their distance from the diagonal. Furrer and Bengtsson [11] proposed to shrink the covariance entries based on this distance from the diagonal. Wu and Pourahmadi [24] and Huang et al. [13] estimate these banded concentration matrices by using thresholding and ℓ_1-penalties respectively, as applied to a Cholesky factor of the inverse covariance matrix. Bickel and Levina [2] prove the consistency of these banded estimators so long as $\frac{(\log\, p)^2}{n} \to 0$ and the model covariance matrix is banded as well, but as they note, these estimators depend on the presented order of the variables.

A related class of models are based on positing some kind of sparsity, either in the covariance matrix, or in the inverse covariance. Bickel and Levina [1] study thresholding estimators of covariance matrices, assuming that each row satisfies an ℓ_q-ball sparsity assumption. In independent work, El Karoui [9] also studied thresholding estimators of the covariance, but based on an alternative notion of sparsity, one which captures the number of closed paths of any length in the associated graph. Other work has studied models in which the inverse covariance or concentration matrix has a sparse structure. As will be clarified in the next section, when the random vector is multivariate Gaussian, the set of non-zero entries in the concentration matrix correspond to the set of edges in an associated Gaussian Markov random field (GMRF). In this setting, imposing sparsity on the concentration matrix can be interpreted as requiring that the graph underlying the GMRF have relatively few edges. A line of recent papers [8, 10, 25] have proposed an estimator that minimizes the Gaussian negative log-likelihood regularized by the ℓ_1 norm of the entries (or the off-diagonal entries) of the concentration matrix. The resulting optimization problem is a log-determinant program, which can be solved in polynomial time with interior point methods [3], or by faster co-ordinate descent algorithms [8, 10]. In recent work, Rothman et al. [21] have analyzed some aspects of high-dimensional behavior of this estimator; assuming that the minimum and maximum eigenvalues of Σ^* are bounded, they show that consistent estimates can be achieved in Frobenius and operator norm, in particular at the rate $\mathcal{O}(\sqrt{\frac{(s+p)\log p}{n}})$.

The focus of this paper is the problem of estimating the concentration matrix Θ^* under sparsity conditions. We do not impose specific distributional assumptions on X itself, but rather analyze the estimator in terms of the tail behavior of the maximum deviation $\max_{i,j} |\widehat{\Sigma}^n_{ij} - \Sigma^*_{ij}|$ of the sample and population covariance matrices. To estimate Θ^*, we consider minimization of an ℓ_1-penalized log-determinant Bregman divergence, which is equivalent to the usual ℓ_1-penalized maximum likelihood when X is multivariate Gaussian. We analyze the behavior of this estimator under high-dimensional scaling, in which the number of nodes p in the graph, and the maximum node degree d are all allowed to grow as a function of the sample size n.

In addition to the triple (n, p, d), we also explicitly keep track of certain other measures of model complexity, that could potentially scale as well. The first of these measures is the ℓ_∞-operator norm of the covariance matrix Σ^*, which we denote by $K_{\Sigma^*} := |||\Sigma^*|||_\infty$. The next quantity involves the Hessian of the log-determinant objective function, $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$. When the distribution of X is multivariate Gaussian, this Hessian has the more explicit representation $\Gamma^*_{(j,k),(\ell,m)} = \mathrm{cov}\{X_j X_k,\ X_\ell X_m\}$, showing that it measures the covariances of the random variables associated with each edge of the graph. For this reason, the matrix Γ^* can be viewed as an edge-based counterpart to the usual node-based covariance matrix Σ^*. Using S to index the variable pairs (i, j) associated with non-zero entries in the inverse covariance. our analysis involves the quantity $K_{\Gamma^*} = |||(\Gamma^*_{SS})^{-1}|||_\infty$. Finally, we also impose a mutual incoherence or irrepresentability condition on the Hessian Γ^*; this condition is similar to assumptions imposed on Σ^* in previous work [22, 26, 19, 23] on the Lasso. We provide some examples where the Lasso irrepresentability condition holds, but our corresponding condition on Γ^* fails; however, we do not know currently whether one condition strictly dominates the other.

Our first result establishes consistency of our estimator $\widehat{\Theta}$ in the elementwise maximum-norm, providing a rate that depends on the tail behavior of the entries in the random matrix $\widehat{\Sigma}^n - \Sigma^*$. For the special case of sub-Gaussian random vectors with concentration matrices having at most d non-zeros per row, a corollary of our analysis is consistency in spectral norm at rate $|||\widehat{\Theta} - \Theta^*|||_2 = \mathcal{O}(\sqrt{(d^2\ \log p)/n})$, with high probability, thereby strengthening previous results [21]. Under the milder restriction of each element of X having bounded $4m$-th moment, the rate in spectral norm is substantially slower—namely, $|||\widehat{\Theta} - \Theta^*|||_2 = \mathcal{O}(d\, p^{1/2m}/\sqrt{n})$—highlighting that the familiar logarithmic dependence on the model size p is linked to particular tail behavior of the distribution of X. Finally, we show that under the same scalings as above, with probability converging to one, the estimate $\widehat{\Theta}$ correctly specifies the zero pattern of

the concentration matrix Θ^*.

The remainder of this paper is organized as follows. In Section 2, we set up the problem and give some background. Section 3 is devoted to statements of our main results, as well as discussion of their consequences. Section 4 provides an outline of the proofs, with the more technical details deferred to appendices. In Section 5, we report the results of some simulation studies that illustrate our theoretical predictions.

Notation For the convenience of the reader, we summarize here notation to be used throughout the paper. Given a vector $u \in \mathbb{R}^d$ and parameter $a \in [1, \infty]$, we use $\|u\|_a$ to denote the usual ℓ_a norm. Given a matrix $U \in \mathbb{R}^{p \times p}$ and parameters $a, b \in [1, \infty]$, we use $|\!|\!| U |\!|\!|_{a,b}$ to denote the induced matrix-operator norm $\max_{\|y\|_a = 1} \|Uy\|_b$; see Horn and Johnson [12] for background. Three cases of particular importance in this paper are the *spectral norm* $|\!|\!| U |\!|\!|_2$, corresponding to the maximal singular value of U; the ℓ_∞/ℓ_∞*-operator norm*, given by

$$|\!|\!| U |\!|\!|_\infty \quad := \quad \max_{j=1,\ldots,p} \sum_{k=1}^{p} |U_{jk}|, \tag{1}$$

and the ℓ_1/ℓ_1*-operator norm*, given by $|\!|\!| U |\!|\!|_1 = |\!|\!| U^T |\!|\!|_\infty$. Finally, we use $\|U\|_\infty$ to denote the element-wise maximum $\max_{i,j} |U_{ij}|$; note that this is not a matrix norm, but rather a norm on the vectorized form of the matrix. For any matrix $U \in \mathbb{R}^{p \times p}$, we use $\mathrm{vec}(U)$ or equivalently $\overline{U} \in \mathbb{R}^{p^2}$ to denote its *vectorized form*, obtained by stacking up the rows of U. We use $\langle\!\langle U, V \rangle\!\rangle := \sum_{i,j} U_{ij} V_{ij}$ to denote the *trace inner product* on the space of symmetric matrices. Note that this inner product induces the *Frobenius norm* $|\!|\!| U |\!|\!|_F := \sqrt{\sum_{i,j} U_{ij}^2}$. Finally, for asymptotics, we use the following standard notation: we write $f(n) = \mathcal{O}(g(n))$ if $f(n) \leq c g(n)$ for some constant $c < \infty$, and $f(n) = \Omega(g(n))$ if $f(n) \geq c' g(n)$ for some constant $c' > 0$. The notation $f(n) \asymp g(n)$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$.

2 Background and problem set-up

Let $X = (X_1, \ldots, X_p)$ be a zero mean p-dimensional random vector. The focus of this paper is the problem of estimating the covariance matrix $\Sigma^* := \mathbb{E}[XX^T]$ and concentration matrix $\Theta^* := \Sigma^{*-1}$ of the random vector X given n i.i.d. observations $\{X^{(k)}\}_{k=1}^n$. In this section, we provide background, and set up this problem more precisely. We begin with background on Gaussian graphical models, which provide one motivation for the estimation of concentration matrices. We then describe an estimator based based on minimizing an ℓ_1 regularized log-determinant divergence; when the data are drawn from a Gaussian graphical model, this estimator corresponds to ℓ_1-regularized maximum likelihood. We then discuss the distributional assumptions that we make in this paper.

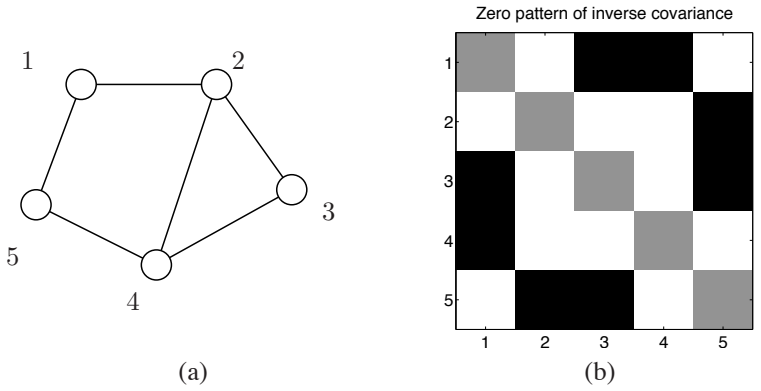
2.1 Gaussian graphical models

One motivation for this paper is the problem of Gaussian graphical model selection. A graphical model or a Markov random field is a family of probability distributions for which the conditional independence and factorization properties are captured by a graph. Let $X = (X_1, X_2, \ldots, X_p)$ denote a zero-mean Gaussian random vector; its density can be parameterized by the inverse covariance or *concentration matrix* $\Theta^* = (\Sigma^*)^{-1} \in \mathcal{S}_+^p$, and can be written as

$$f(x_1, \ldots, x_p; \Theta^*) \quad = \quad \frac{1}{\sqrt{(2\pi)^p \det((\Theta^*)^{-1})}} \exp\Big\{ -\frac{1}{2} x^T \Theta^* x \Big\}. \tag{2}$$

We can relate this Gaussian distribution of the random vector X to a graphical model as follows. Suppose we are given an undirected graph $G = (V, E)$ with vertex set $V = \{1, 2, \ldots, p\}$ and edge[1] set E, so that each variable X_i is associated with a corresponding vertex $i \in V$. The Gaussian Markov random field (GMRF) associated with the graph

[1] As a remark on notation, we would like to contrast the notation for the edge-set E from the notation for an expectation of a random variable, $\mathbb{E}(\cdot)$.



(a) (b)

Figure 1. (a) Simple undirected graph. A Gauss Markov random field has a Gaussian variable X_i associated with each vertex $i \in V$. This graph has $p = 5$ vertices, maximum degree $d = 3$ and $s = 6$ edges. (b) Zero pattern of the inverse covariance Θ^* associated with the GMRF in (a). The set $E(\Theta^*)$ corresponds to the off-diagonal non-zeros (white blocks); the diagonal is also non-zero (grey squares), but these entries do not correspond to edges. The black squares correspond to non-edges, or zeros in Θ^*.

G over the random vector X is then the family of Gaussian distributions with concentration matrices Θ^* that respect the edge structure of the graph, in the sense that $\Theta^*_{ij} = 0$ if $(i, j) \notin E$. Figure 1 illustrates this correspondence between the graph structure (panel (a)), and the sparsity pattern of the concentration matrix Θ^* (panel (b)). The problem of estimating the entries of the concentration matrix Θ^* corresponds to estimating the Gaussian graphical model instance, while the problem of estimating the off-diagonal zero-pattern of the concentration matrix—-that is, the set

$$E(\Theta^*) \quad := \quad \{i, j \in V \mid i \neq j, \Theta^*_{ij} \neq 0\} \tag{3}$$

corresponds to the problem of Gaussian graphical *model selection*.

With a slight abuse of notation, we define the *sparsity index* $s := |E(\Theta^*)|$ as the total number of non-zero elements in off-diagonal positions of Θ^*; equivalently, this corresponds to twice the number of edges in the case of a Gaussian graphical model. We also define the *maximum degree or row cardinality*

$$d \quad := \quad \max_{i=1,\ldots,p} \Big| \big\{ j \in V \ \mid \ \Theta^*_{ij} \neq 0 \big\} \Big|, \tag{4}$$

corresponding to the maximum number of non-zeros in any row of Θ^*; this corresponds to the maximum degree in the graph of the underlying Gaussian graphical model. Note that we have included the diagonal entry Θ^*_{ii} in the degree count, corresponding to a self-loop at each vertex.

It is convenient throughout the paper to use graphical terminology, such as degrees and edges, even though the distributional assumptions that we impose, as described in Section 2.3, are milder and hence apply even to distributions that are not Gaussian MRFs.

2.2 ℓ_1-penalized log-determinant divergence

An important set in this paper is the cone

$$\mathcal{S}^p_+ \quad := \quad \big\{ A \in \mathbb{R}^{p \times p} \mid A = A^T, \ A \succeq 0 \big\}, \tag{5}$$

formed by all symmetric positive semi-definite matrices in p dimensions. We assume that the covariance matrix Σ^* and concentration matrix Θ^* of the random vector X are strictly positive definite, and so lie in the interior of this cone $\mathcal{S}^p_+$.

The focus of this paper is a particular type of M-estimator for the concentration matrix Θ^*, based on minimizing a Bregman divergence between symmetric matrices. A function is of Bregman type if it is strictly convex, continuously differentiable and has bounded level sets [4, 7]. Any such function induces a *Bregman divergence* of the form

$D_g(A\|B) = g(A) - g(B) - \langle \nabla g(B), A - B \rangle$. From the strict convexity of g, it follows that $D_g(A\|B) \geq 0$ for all A and B, with equality if and only if $A = B$.

As a candidate Bregman function, consider the log-determinant barrier function, defined for any matrix $A \in \mathcal{S}_+^p$ by

$$g(A) \quad := \quad \begin{cases} -\log\det(A) & \text{if } A \succ 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{6}$$

As is standard in convex analysis, we view this function as taking values in the extended reals $\mathbb{R}_* = \mathbb{R} \cup \{+\infty\}$. With this definition, the function g is strictly convex, and its domain is the set of strictly positive definite matrices. Moreover, it is continuously differentiable over its domain, with $\nabla g(A) = -A^{-1}$; see Boyd and Vandenberghe [3] for further discussion. The Bregman divergence corresponding to this log-determinant Bregman function g is given by

$$D_g(A\|B) \quad := \quad -\log\det A + \log\det B + \langle\langle B^{-1},\ A - B \rangle\rangle, \tag{7}$$

valid for any $A, B \in \mathcal{S}_+^p$ that are strictly positive definite. This divergence suggests a natural way to estimate concentration matrices—namely, by minimizing the divergence $D_g(\Theta^*\|\Theta)$—or equivalently, by minimizing the function

$$\min_{\Theta \succ 0} \big\{ \langle\langle \Theta,\ \Sigma^* \rangle\rangle - \log\det\Theta \big\}, \tag{8}$$

where we have discarded terms independent of Θ, and used the fact that the inverse of the concentration matrix is the covariance matrix (i.e., $(\Theta^*)^{-1} = \Sigma^* = \mathbb{E}[XX^T]$). Of course, the convex program (8) cannot be solved without knowledge of the true covariance matrix Σ^*, but one can take the standard approach of replacing Σ^* with an empirical version, with the possible addition of a regularization term.

In this paper, we analyze a particular instantiation of this strategy. Given n samples, we define the *sample covariance matrix*

$$\widehat{\Sigma}^n \quad := \quad \frac{1}{n} \sum_{k=1}^{n} X^{(k)} (X^{(k)})^T. \tag{9}$$

To lighten notation, we occasionally drop the superscript n, and simply write $\widehat{\Sigma}$ for the sample covariance. We also define the *off-diagonal ℓ_1 regularizer*

$$\|\Theta\|_{1,\text{off}} \quad := \quad \sum_{i \neq j} |\Theta_{ij}|, \tag{10}$$

where the sum ranges over all $i, j = 1, \ldots, p$ with $i \neq j$. Given some regularization constant $\lambda_n > 0$, we consider estimating Θ^* by solving the following *ℓ_1-regularized log-determinant program*:

$$\widehat{\Theta} \quad := \quad \arg\min_{\Theta \succ 0} \big\{ \langle\langle \Theta,\ \widehat{\Sigma}^n \rangle\rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \big\}. \tag{11}$$

As shown in Appendix A, for any $\lambda_n > 0$ and sample covariance matrix $\widehat{\Sigma}^n$ with strictly positive diagonal, this convex optimization problem has a unique optimum, so there is no ambiguity in equation (11). When the data is actually drawn from a multivariate Gaussian distribution, then the problem (11) is simply ℓ_1-regularized maximum likelihood.

2.3 Tail conditions

In this section, we describe the tail conditions that underlie our analysis. Since the estimator (11) is based on using the sample covariance $\widehat{\Sigma}^n$ as a surrogate for the (unknown) covariance Σ^*, any type of consistency requires bounds on the difference $\widehat{\Sigma}^n - \Sigma^*$. In particular, we define the following tail condition:

Definition 1 (Tail conditions)**.** The random vector X satisfies tail condition $\mathcal{T}(f, v_*)$ if there exists a constant $v_* \in (0, \infty]$ and a function $f : \mathbb{N} \times (0, \infty) \to (0, \infty)$ such that for any $(i, j) \in V \times V$:

$$\mathbb{P}[|\widehat{\Sigma}^n_{ij} - \Sigma^*_{ij}| \geq \delta] \quad \leq \quad 1/f(n, \delta) \qquad \text{for all } \delta \in (0, 1/v_*]. \tag{12}$$

We adopt the convention $1/0 := +\infty$, so that the value $v_* = 0$ indicates the inequality holds for any $\delta \in (0, \infty)$.

Two important examples of the tail function f are the following:

(a) an *exponential-type tail function*, meaning that $f(n, \delta) = \exp(c\, n\, \delta^a)$, for some scalar $c > 0$, and exponent $a > 0$; and

(b) a *polynomial-type tail function*, meaning that $f(n, \delta) = c\, n^m\, \delta^{2m}$, for some positive integer $m \in \mathbb{N}$ and scalar $c > 0$.

As might be expected, if X is multivariate Gaussian, then the deviations of sample covariance matrix have an exponential-type tail function with $a = 2$. A bit more generally, in the following subsections, we provide broader classes of distributions whose sample covariance entries satisfy exponential and a polynomial tail bounds (see Lemmata 1 and 2 respectively).

Given a larger number of samples n, we expect the tail probability bound $1/f(n, \delta)$ to be smaller, or equivalently, for the tail function $f(n, \delta)$ to larger. Accordingly, we require that f is monotonically increasing in n, so that for each fixed $\delta > 0$, we can define the inverse function

$$\overline{n}_f(r; \delta) \quad := \quad \arg\max \big\{ n \;\; | \;\; f(n, \delta) \leq r \big\}. \tag{13}$$

Similarly, we expect that f is monotonically increasing in δ, so that for each fixed n, we can define the inverse in the second argument

$$\bar{\delta}_f(r; n) \quad := \quad \arg\max \big\{ \delta \;\; | \;\; f(n, \delta) \leq r \big\}. \tag{14}$$

For future reference, we note a simple consequence of the monotonicity of the tail function f—namely

$$n > \overline{n}_f(\delta, r) \quad \text{for some } \delta > 0 \quad \Longrightarrow \quad \bar{\delta}_f(n, r) \leq \delta. \tag{15}$$

The inverse functions $\overline{n}_f$ and $\bar{\delta}_f$ play an important role in describing the behavior of our estimator. We provide concrete examples in the following two subsections.

2.3.1 Sub-Gaussian distributions

In this subsection, we study the case of i.i.d. observations of sub-Gaussian random variables.

Definition 2. A zero-mean random variable Z is *sub-Gaussian* if there exists a constant $\sigma \in (0, \infty)$ such that

$$\mathbb{E}[\exp(tZ)] \quad \leq \quad \exp(\sigma^2\, t^2/2) \qquad \text{for all } t \in \mathbb{R}. \tag{16}$$

By the Chernoff bound, this upper bound (16) on the moment-generating function implies a two-sided tail bound of the form

$$\mathbb{P}[|Z| > z] \quad \leq \quad 2 \exp\big(- \frac{z^2}{2\sigma^2} \big). \tag{17}$$

Naturally, any zero-mean Gaussian variable with variance σ^2 satisfies the bounds (16) and (17). In addition to the Gaussian case, the class of sub-Gaussian variates includes any bounded random variable (e.g., Bernoulli, multinomial, uniform), any random variable with strictly log-concave density [6, 17], and any finite mixture of sub-Gaussian variables.

The following lemma, proved in Appendix D, shows that the entries of the sample covariance based on i.i.d. samples of sub-Gaussian random vector satisfy an exponential-type tail bound with exponent $a = 2$. The argument is along the lines of a result due to Bickel and Levina [1], but with more explicit control of the constants in the error exponent:

Lemma 1. *Consider a zero-mean random vector $(X_1, \ldots, X_p)$ with covariance Σ^* such that each $X_i / \sqrt{\Sigma^*_{ii}}$ is sub-Gaussian with parameter σ. Given n i.i.d. samples, the associated sample covariance $\widehat{\Sigma}^n$ satisfies the tail bound*

$$\mathbb{P}\big[|\widehat{\Sigma}^n_{ij} - \Sigma^*_{ij}| > \delta \big] \quad \leq \quad 4 \exp\Big\{ - \frac{n\delta^2}{128(1 + 4\sigma^2)^2 \max_i (\Sigma^*_{ii})^2} \Big\},$$

for all $\delta \in \big(0, \max_i(\Sigma^*_{ii})\, 8(1 + 4\sigma^2)\big)$.

Thus, the sample covariance entries the tail condition $\mathcal{T}(f, v_*)$ with $v_* = \left[\max_i(\Sigma^*_{ii})\, 8(1+4\sigma^2)\right]^{-1}$, and an exponential-type tail function with $a = 2$—namely

$$f(n,\delta) = \frac{1}{4}\exp(c_* n\delta^2), \quad \text{with} \quad c_* = \left[128(1+4\sigma^2)^2 \max_i(\Sigma^*_{ii})^2\right]^{-1} \tag{18}$$

A little calculation shows that the associated inverse functions take the form

$$\bar{\delta}_f(r;n) = \sqrt{\frac{\log(4\,r)}{c_*\, n}}, \quad \text{and} \quad \overline{n}_f(r;\delta) = \frac{\log(4\,r)}{c_*\delta^2}. \tag{19}$$

2.3.2 Tail bounds with moment bounds

In the following lemma, proved in Appendix E, we show that given i.i.d. observations from random variables with bounded moments, the sample covariance entries satisfy a polynomial-type tail bound. See the papers [26, 9] for related results on tail bounds for variables with bounded moments.

Lemma 2. *Suppose there exists a positive integer m and scalar $K_m \in \mathbb{R}$ such that for $i = 1, \ldots, p$,*

$$\mathbb{E}\left[\left(\frac{X_i}{\sqrt{\Sigma^*_{ii}}}\right)^{4m}\right] \leq K_m. \tag{20}$$

For i.i.d. samples $\{X_i^{(k)}\}_{k=1}^n$, the sample covariance matrix $\widehat{\Sigma}^n$ satisfies the bound

$$\mathbb{P}\left[\left|\widehat{\Sigma}^n_{ij} - \Sigma^*_{ij}\right| > \delta\right] \leq \frac{\left\{m^{2m+1}2^{2m}(\max_i \Sigma^*_{ii})^{2m}\,(K_m+1)\right\}}{n^m\,\delta^{2m}}. \tag{21}$$

Thus, in this case, the sample covariance satisfies the tail condition $\mathcal{T}(f, v_*)$ with $v_* = 0$, so that the bound holds for all $\delta \in (0, \infty)$, and with the polynomial-type tail function

$$f(n,\delta) = c_* n^m \delta^{2m} \quad \text{where } c_* = 1/\left\{m^{2m+1}2^{2m}(\max_i \Sigma^*_{ii})^{2m}\,(K_m+1)\right\}. \tag{22}$$

Finally, a little calculation shows that in this case, the inverse tail functions take the form

$$\bar{\delta}_f(n,r) = \frac{(r/c_*)^{1/2m}}{\sqrt{n}}, \quad \text{and} \quad \overline{n}_f(\delta,r) = \frac{(r/c_*)^{1/m}}{\delta^2}. \tag{23}$$

3 Main results and some consequences

In this section, we state our main results, and discuss some of their consequences. We begin in Section 3.1 by stating some conditions on the true concentration matrix Θ^* required in our analysis, including a particular type of incoherence or irrepresentability condition. In Section 3.2, we state our first main result—namely, Theorem 1 on consistency of the estimator $\widehat{\Theta}$, and the rate of decay of its error in elementwise ℓ_∞ norm. Section 3.3 is devoted to Theorem 2 on the model selection consistency of the estimator. Section 3.4 is devoted the relation between the log-determinant estimator and the ordinary Lasso (neighborhood-based approach) as methods for graphical model selection; in addition, we illustrate our irrepresentability assumption for some simple graphs. Finally, in Section 3.5, we state and prove some corollaries of Theorem 1, regarding rates in Frobenius and operator norms.

3.1 Conditions on covariance and Hessian

Our results involve some quantities involving the Hessian of the log-determinant barrier (6), evaluated at the true concentration matrix Θ^*. Using standard results on matrix derivatives [3], it can be shown that this Hessian takes the form

$$\Gamma^* \;:=\; \nabla^2_\Theta g(\Theta)\Big|_{\Theta=\Theta^*} \;=\; \Theta^{*-1} \otimes \Theta^{*-1}, \tag{24}$$

where $\otimes$ denotes the Kronecker matrix product. By definition, Γ^* is a $p^2 \times p^2$ matrix indexed by vertex pairs, so that entry $\Gamma^*_{(j,k),(\ell,m)}$ corresponds to the second partial derivative $\frac{\partial^2 g}{\partial \Theta_{jk} \partial \Theta_{\ell m}}$, evaluated at $\Theta = \Theta^*$. When X has multivariate Gaussian distribution, then Γ^* is the Fisher information of the model, and by standard results on cumulant functions in exponential families [5], we have the more specific expression $\Gamma^*_{(j,k),(\ell,m)} = \mathrm{cov}\{X_j X_k, \ X_\ell X_m\}$. For this reason, Γ^* can be viewed as an edge-based counterpart to the usual covariance matrix Σ^*.

We define the set of non-zero off-diagonal entries in the model concentration matrix Θ^*:

$$E(\Theta^*) \quad := \quad \{(i,j) \in V \times V \ \mid \ i \neq j, \Theta^*_{ij} \neq 0\}, \tag{25}$$

and let $S(\Theta^*) = \{E(\Theta^*) \cup \{(1,1), \ldots, (p,p)\}$ be the augmented set including the diagonal. We let $S^c(\Theta^*)$ denote the complement of $S(\Theta^*)$ in the set $\{1, \ldots, p\} \times \{1, \ldots, p\}$, corresponding to all pairs (ℓ, m) for which $\Theta^*_{\ell m} = 0$. When it is clear from context, we shorten our notation for these sets to S and S^c, respectively. Finally, for any two subsets T and T' of $V \times V$, we use $\Gamma^*_{TT'}$ to denote the $|T| \times |T'|$ matrix with rows and columns of Γ^* indexed by T and T' respectively.

Our main results involve the ℓ_∞/ℓ_∞ norm applied to the covariance matrix Σ^*, and to the inverse of a sub-block of the Hessian Γ^*. In particular, we define

$$K_{\Sigma^*} \quad := \quad |\!|\!|\Sigma^*|\!|\!|_\infty \ = \ \Big(\max_{i=1,\ldots,p} \sum_{j=1}^{p} |\Sigma^*_{ij}| \Big), \tag{26}$$

corresponding to the ℓ_∞-operator norm of the true covariance matrix Σ^*, and

$$K_{\Gamma^*} \quad := \quad |\!|\!|(\Gamma^*_{SS})^{-1}|\!|\!|_\infty \ = \ |\!|\!|([{\Theta^*}^{-1} \otimes {\Theta^*}^{-1}]_{SS})^{-1}|\!|\!|_\infty. \tag{27}$$

Our analysis keeps explicit track of these quantities, so that they can scale in a non-trivial manner with the problem dimension p.

We assume the Hessian satisfies the following type of *mutual incoherence or irrepresentable condition*:

Assumption 1. There exists some $\alpha \in (0, 1]$ such that

$$|\!|\!|\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}|\!|\!|_\infty \quad \leq \quad (1 - \alpha). \tag{28}$$

The underlying intuition is that this assumption imposes control on the influence that the non-edge terms, indexed by S^c, can have on the edge-based terms, indexed by S. It is worth noting that a similar condition for the Lasso, with the covariance matrix Σ^* taking the place of the matrix Γ^* above, is necessary and sufficient for support recovery using the ordinary Lasso [19, 22, 23, 26]. See Section 3.4 for illustration of the form taken by Assumption 1 for specific graphical models.

A remark on notation: although our analysis allows the quantities $K_{\Sigma^*}, K_{\Gamma^*}$ as well as the model size p and maximum node-degree d to grow with the sample size n, we suppress this dependence on n in their notation.

3.2 Rates in elementwise ℓ_∞-norm

We begin with a result that provides sufficient conditions on the sample size n for bounds in the elementwise ℓ_∞-norm. This result is stated in terms of the tail function f, and its inverses $\overline{n}_f$ and $\bar{\delta}_f$ (equations (13) and (14)), and so covers a general range of possible tail behaviors. So as to make it more concrete, we follow the general statement with corollaries for the special cases of exponential-type and polynomial-type tail functions, corresponding to sub-Gaussian and moment-bounded variables respectively.

In the theorem statement, the choice of regularization constant λ_n is specified in terms of a user-defined parameter $\tau > 2$. Larger choices of τ yield faster rates of convergence in the probability with which the claims hold, but also lead to more stringent requirements on the sample size.

Theorem 1. *Consider a distribution satisfying the incoherence assumption (28) with parameter $\alpha \in (0,1]$, and the tail condition (12) with parameters $\mathcal{T}(f, v_*)$. Let $\widehat{\Theta}$ be the unique optimum of the log-determinant program (11) with regularization parameter $\lambda_n = (8/\alpha)\,\bar{\delta}_f(n, p^\tau)$ for some $\tau > 2$. Then, if the sample size is lower bounded as*

$$n \;>\; \bar{n}_f\left(1\Big/ \max\Big\{v_*,\; 6\big(1+8\alpha^{-1}\big)\, d\, \max\{K_{\Sigma^*}K_{\Gamma^*}, K_{\Sigma^*}^3 K_{\Gamma^*}^2\}\Big\},\;\; p^\tau\right), \tag{29}$$

then with probability greater than $1 - 1/p^{\tau-2} \to 1$, we have:

(a) The estimate $\widehat{\Theta}$ satisfies the elementwise ℓ_∞-bound:

$$\|\widehat{\Theta} - \Theta^*\|_\infty \;\;\leq\;\; \big\{2\big(1+8\alpha^{-1}\big)K_{\Gamma^*}\big\}\;\bar{\delta}_f(n, p^\tau). \tag{30}$$

(b) It specifies an edge set $E(\widehat{\Theta})$ that is a subset of the true edge set $E(\Theta^*)$, and includes all edges (i,j) with $|\Theta^*_{ij}| > \big\{2\big(1+8\alpha^{-1}\big)K_{\Gamma^*}\big\}\;\bar{\delta}_f(n, p^\tau)$.

If we assume that the various quantities $K_{\Gamma^*}, K_{\Sigma^*}, \alpha$ remain constant as a function of (n, p, d), we have the elementwise ℓ_∞ bound $\|\widehat{\Theta} - \Theta^*\|_\infty = \mathcal{O}(\bar{\delta}_f(n, p^\tau))$, so that the inverse tail function $\bar{\delta}_f(n, p^\tau)$ (see equation (14)) specifies rate of convergence in the element-wise ℓ_∞-norm. In the following section, we derive the consequences of this ℓ_∞-bound for two specific tail functions, namely those of exponential-type with $a = 2$, and polynomial-type tails (see Section 2.3). Turning to the other factors involved in the theorem statement, the quantities K_{Σ^*} and K_{Γ^*} measure the sizes of the entries in the covariance matrix Σ^* and inverse Hessian $(\Gamma^*)^{-1}$ respectively. Finally, the factor $(1+\frac{8}{\alpha})$ depends on the irrepresentability assumption 1, growing in particular as the incoherence parameter α approaches 0.

3.2.1 Exponential-type tails

We now discuss the consequences of Theorem 1 for distributions in which the sample covariance satisfies an exponential-type tail bound with exponent $a = 2$. In particular, recall from Lemma 1 that such a tail bound holds when the variables are sub-Gaussian.

Corollary 1. *Under the same conditions as Theorem 1, suppose moreover that the variables $X_i/\sqrt{\Sigma^*_{ii}}$ are sub-Gaussian with parameter σ, and the samples are drawn independently. Then if the sample size n satisfies the bound*

$$n \;>\; C_1\; d^2\,(1+\frac{8}{\alpha})^2\;\big(\tau \log p + \log 4\big) \tag{31}$$

where $C_1 := \big\{48\sqrt{2}\,(1+4\sigma^2)\,\max_i(\Sigma^*_{ii})\,\max\{K_{\Sigma^*}K_{\Gamma^*}, K_{\Sigma^*}^3K_{\Gamma^*}^2\}\big\}^2$, then with probability greater than $1 - 1/p^{\tau-2}$, the estimate $\widehat{\Theta}$ satisfies the bound,

$$\|\widehat{\Theta} - \Theta^*\|_\infty \;\;\leq\;\; \big\{16\sqrt{2}\,(1+4\sigma^2)\;\max_i(\Sigma^*_{ii})\,(1+8\alpha^{-1})K_{\Gamma^*}\big\}\;\sqrt{\frac{\tau\log p + \log 4}{n}}.$$

Proof. From Lemma 1, when the rescaled variables $X_i/\sqrt{\Sigma^*_{ii}}$ are sub-Gaussian with parameter σ, the sample covariance entries satisfies a tail bound $\mathcal{T}(f, v_*)$ with with $v_* = \big[\max_i(\Sigma^*_{ii})\,8(1+4\sigma^2)\big]^{-1}$ and $f(n,\delta) = (1/4)\exp(c_* n\delta^2)$, where $c_* = \big[128(1+4\sigma^2)^2\max_i(\Sigma^*_{ii})^2\big]^{-1}$. As a consequence, for this particular model, the inverse functions $\bar{\delta}_f(n, p^\tau)$ and $\bar{n}_f(\delta, p^\tau)$ take the form

$$\bar{\delta}_f(n, p^\tau) \;\;=\;\; \sqrt{\frac{\log(4\,p^\tau)}{c_*\,n}} \;\;=\;\; \sqrt{128(1+4\sigma^2)^2\max_i(\Sigma^*_{ii})^2}\;\sqrt{\frac{\tau\log p + \log 4}{n}}, \tag{32a}$$

$$\bar{n}_f(\delta, p^\tau) \;\;=\;\; \frac{\log(4\,p^\tau)}{c_*\delta^2} \;\;=\;\; 128(1+4\sigma^2)^2\max_i(\Sigma^*_{ii})^2\;\left(\frac{\tau\log p + \log 4}{\delta^2}\right). \tag{32b}$$

Substituting these forms into the claim of Theorem 1 and doing some simple algebra yields the stated corollary. □

When $K_{\Gamma^*}, K_{\Sigma^*}, \alpha$ remain constant as a function of (n, p, d), the corollary can be summarized succinctly as a sample size of $n = \Omega(d^2 \log p)$ samples ensures that an elementwise ℓ_∞ bound $\|\widehat{\Theta} - \Theta^*\|_\infty = \mathcal{O}(\sqrt{\frac{\log p}{n}})$ holds with high probability. In practice, one frequently considers graphs with maximum node degrees d that either remain bounded, or that grow sub-linearly with the graph size (i.e., $d = o(p)$). In such cases, the sample size allowed by the corollary can be substantially smaller than the graph size, so that for sub-Gaussian random variables, the method can succeed in the $p \gg n$ regime.

3.2.2 Polynomial-type tails

We now state a corollary for the case of a polynomial-type tail function, such as those ensured by the case of random variables with appropriately bounded moments.

Corollary 2. *Under the assumptions of Theorem 1, suppose the rescaled variables $X_i/\sqrt{\Sigma^*_{ii}}$ have $4m^{th}$ moments upper bounded by K_m, and the sampling is i.i.d. Then if the sample size n satisfies the bound*

$$n \quad > \quad C_2\, d^2 \left(1 + \frac{8}{\alpha}\right)^2 p^{\tau/m}, \tag{33}$$

where $C_2 := \left\{12m\,[m(K_m+1)]^{\frac{1}{2m}}\, \max_i(\Sigma^*_{ii}) \max\{K^2_{\Sigma^*} K_{\Gamma^*}, K^4_{\Sigma^*} K^2_{\Gamma^*}\}\right\}^2$, then with probability greater than $1 - 1/p^{\tau-2}$, the estimate $\widehat{\Theta}$ satisfies the bound,

$$\|\widehat{\Theta} - \Theta^*\|_\infty \quad \leq \quad \{4m[m(K_m+1)]^{\frac{1}{2m}} \left(1 + \frac{8}{\alpha}\right) K_{\Gamma^*}\} \sqrt{\frac{p^{\tau/m}}{n}}.$$

Proof. Recall from Lemma 2 that when the rescaled variables $X_i/\sqrt{\Sigma^*_{ii}}$ have bounded $4m^{th}$ moments, then the sample covariance $\widehat{\Sigma}$ satisfies the tail condition $\mathcal{T}(f, v_*)$ with $v_* = 0$, and with $f(n, \delta) = c_* n^m \delta^{2m}$ with c_* defined as $c_* = 1/\left\{m^{2m+1} 2^{2m} (\max_i \Sigma^*_{ii})^{2m}\, (K_m + 1)\right\}$. As a consequence, for this particular model, the inverse functions take the form

$$\bar{\delta}_f(n, p^\tau) \quad = \quad \frac{(p^\tau/c_*)^{1/2m}}{\sqrt{n}} \quad = \quad \{2m[m(K_m+1)]^{\frac{1}{2m}} \max_i \Sigma^*_{ii}\} \sqrt{\frac{p^{\tau/m}}{n}}, \tag{34a}$$

$$\bar{n}_f(\delta, p^\tau) \quad = \quad \frac{(p^\tau/c_*)^{1/m}}{\delta^2} \quad = \quad \{2m[m(K_m+1)]^{\frac{1}{2m}} \max_i \Sigma^*_{ii}\}^2 \left(\frac{p^{\tau/m}}{\delta^2}\right). \tag{34b}$$

The claim then follows by substituting these expressions into Theorem 1 and performing some algebra. □

When the quantities $(K_{\Gamma^*}, K_{\Sigma^*}, \alpha)$ remain constant as a function of (n, p, d), Corollary 2 can be summarized succinctly as $n = \Omega(d^2\, p^{\tau/m})$ samples are sufficient to achieve a convergence rate in elementwise ℓ_∞-norm of the order $\|\widehat{\Theta} - \Theta^*\|_\infty = \mathcal{O}(\sqrt{\frac{p^{\tau/m}}{n}})$, with high probability. Consequently, both the required sample size and the rate of convergence of the estimator are polynomial in the number of variables p. It is worth contrasting these rates with the case of sub-Gaussian random variables, where the rates have only logarithmic dependence on the problem size p.

3.3 Model selection consistency

Part (b) of Theorem 1 asserts that the edge set $E(\widehat{\Theta})$ returned by the estimator is contained within the true edge set $E(\Theta^*)$—meaning that it correctly *excludes* all non-edges—and that it includes all edges that are "large", relative to the $\bar{\delta}_f(n, p^\tau)$ decay of the error. The following result, essentially a minor refinement of Theorem 1, provides sufficient conditions linking the sample size n and the minimum value

$$\theta_{\min} \quad := \quad \min_{(i,j) \in E(\Theta^*)} |\Theta^*_{ij}| \tag{35}$$

for model selection consistency. More precisely, define the event

$$\mathcal{M}(\widehat{\Theta}; \Theta^*) \quad := \quad \left\{\text{sign}(\widehat{\Theta}_{ij}) = \text{sign}(\Theta^*_{ij}) \quad \forall (i,j) \in E(\Theta^*)\right\} \tag{36}$$

that the estimator $\widehat{\Theta}$ has the same edge set as Θ^*, and moreover recovers the correct signs on these edges. With this notation, we have:

Theorem 2. *Under the same conditions as Theorem 1, suppose that the sample size satisfies the lower bound*

$$n \quad > \quad \overline{n}_f\left(1/\max\left\{2K_{\Gamma^*}(1+8\alpha^{-1})\,\theta_{\min}^{-1},\; v_*,\; 6\big(1+8\alpha^{-1}\big)\,d\,\max\{K_{\Sigma^*}K_{\Gamma^*},K_{\Sigma^*}^3K_{\Gamma^*}^2\}\right\},\;\; p^\tau\right). \quad (37)$$

Then the estimator is model selection consistent with high probability as $p \to \infty$,

$$\mathbb{P}\big[\mathcal{M}(\widehat{\Theta};\Theta^*)\big] \quad \geq \quad 1 - 1/p^{\tau-2} \;\to\; 1. \quad (38)$$

In comparison to Theorem 1, the sample size requirement (37) differs only in the additional term $\frac{2K_{\Gamma^*}(1+\frac{8}{\alpha})}{\theta_{\min}}$ involving the minimum value. This term can be viewed as constraining how quickly the minimum can decay as a function of (n,p), as we illustrate with some concrete tail functions.

3.3.1 Exponential-type tails

Recall the setting of Section 2.3.1, where the random variables $\{X_i^{(k)}/\sqrt{\Sigma^*_{ii}}\}$ are sub-Gaussian with parameter σ. Let us suppose that the parameters $(K_{\Gamma^*}, K_{\Sigma^*}, \alpha)$ are viewed as constants (not scaling with (p,d). Then, using the expression (32) for the inverse function $\overline{n}_f$ in this setting, a corollary of Theorem 2 is that a sample size

$$n \quad = \quad \Omega\big((d^2+\theta_{\min}^{-2})\,\tau\log p\big) \quad (39)$$

is sufficient for model selection consistency with probability greater than $1-1/p^{\tau-2}$. Alternatively, we can state that $n = \Omega(\tau d^2\log p)$ samples are sufficient, as along as the minimum value scales as $\theta_{\min} = \Omega(\sqrt{\frac{\log p}{n}})$.

3.3.2 Polynomial-type tails

Recall the setting of Section 2.3.2, where the rescaled random variables $X_i/\sqrt{\Sigma^*_{ii}}$ have bounded $4m^{th}$ moments. Using the expression (34) for the inverse function $\overline{n}_f$ in this setting, a corollary of Theorem 2 is that a sample size

$$n \quad = \quad \Omega\big((d^2+\theta_{\min}^{-2})\,p^{\tau/m}\big) \quad (40)$$

is sufficient for model selection consistency with probability greater than $1-1/p^{\tau-2}$. Alternatively, we can state than $n = \Omega(d^2p^{\tau/m})$ samples are sufficient, as long as the minimum value scales as $\theta_{\min} = \Omega(p^{\tau/(2m)}/\sqrt{n})$.

3.4 Comparison to neighbor-based graphical model selection

Suppose that X follows a multivariate Gaussian distribution, so that the structure of the concentration matrix Θ^* specifies the structure of a Gaussian graphical model. In this case, it is interesting to compare our sufficient conditions for graphical model consistency of the log-determinant approach, as specified in Theorem 2, to those of the neighborhood-based method, first proposed by Meinshausen and Bühlmann [19]. The latter method estimates the full graph structure by performing an ℓ_1-regularized linear regression (Lasso)—of the form $X_i = \sum_{j\neq i}\theta_{ij}X_j + W$— of each node on its neighbors and using the support of the estimated regression vector θ to predict the neighborhood set. These neighborhoods are then combined, by either an OR rule or an AND rule, to estimate the full graph. Various aspects of the high-dimensional model selection consistency of the Lasso are now understood [19, 23, 26]; for instance, it is known that mutual incoherence or irrepresentability conditions are necessary and sufficient for its success [22, 26]. In terms of scaling, Wainwright [23] shows that the Lasso succeeds with high probability if and only if the sample size scales as $n \asymp c(\{d+\theta_{\min}^{-2}\}\log p)$, where c is a constant determined by the covariance matrix Σ^*. By a union bound over the p nodes in the graph, it then follows that the neighbor-based graph selection method in turn succeeds with high probability if $n = \Omega(\{d+\theta_{\min}^{-2}\}\log p)$.

For comparison, consider the application of Theorem 2 to the case where the variables are sub-Gaussian (which includes the Gaussian case). For this setting, we have seen that the scaling required by Theorem 2 is $n = \Omega(\{d^2 + \theta_{\min}^{-2}\} \log p)$, so that the dependence of the log-determinant approach in $\theta_{\min}$ is identical, but it depends quadratically on the maximum degree d. We suspect that that the quadratic dependence d^2 might be an artifact of our analysis, but have not yet been able to reduce it to d. Otherwise, the primary difference between the two methods is in the nature of the irrepresentability assumptions that are imposed: our method requires Assumption 1 on the Hessian Γ^*, whereas the neighborhood-based method imposes this same type of condition on a set of p covariance matrices, each of size $(p-1) \times (p-1)$, one for each node of the graph. Below we show two cases where the Lasso irrepresentability condition holds, while the log-determinant requirement fails. However, in general, we do not know whether the log-determinant irrepresentability strictly dominates its analog for the Lasso.

3.4.1 Illustration of irrepresentability: Diamond graph

Consider the following Gaussian graphical model example from Meinshausen [18]. Figure 2(a) shows a diamond-shaped graph $G = (V, E)$, with vertex set $V = \{1, 2, 3, 4\}$ and edge-set as the fully connected graph over V with the edge $(1, 4)$ removed. The covariance matrix Σ^* is parameterized by the correlation parameter $\rho \in [0, 1/\sqrt{2}]$:

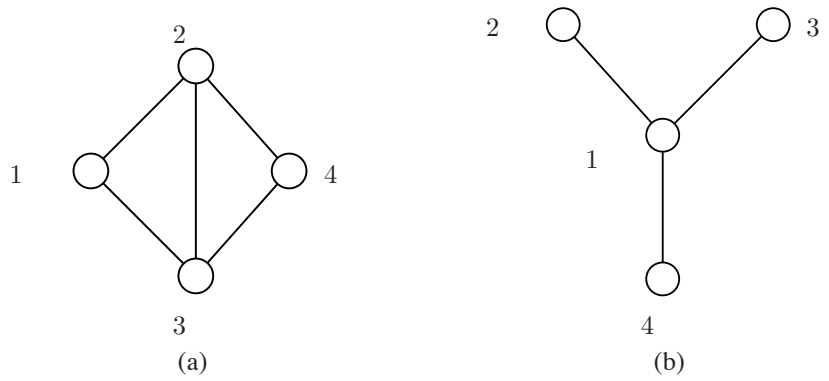


Figure 2: (a) Graph of the example discussed by Meinshausen [18]. (b) A simple 4-node star graph.

the diagonal entries are set to $\Sigma^*_i = 1$, for all $i \in V$; the entries corresponding to edges are set to $\Sigma^*_{ij} = \rho$ for $(i, j) \in E \backslash \{(2, 3)\}$, $\Sigma^*_{23} = 0$; and finally the entry corresponding to the non-edge is set as $\Sigma^*_{14} = 2\rho^2$. Meinshausen [18] showed that the ℓ_1-penalized log-determinant estimator $\widehat{\Theta}$ fails to recover the graph structure, for any sample size, if $\rho > -1 + (3/2)^{1/2} \approx 0.23$. It is instructive to compare this necessary condition to the sufficient condition provided in our analysis, namely the incoherence Assumption 1 as applied to the Hessian Γ^*. For this particular example, a little calculation shows that Assumption 1 is equivalent to the constraint

$$4|\rho|(|\rho| + 1) \quad < \quad 1,$$

an inequality which holds for all $\rho \in (-0.2017, 0.2017)$. Note that the upper value 0.2017 is just below the necessary threshold discussed by Meinshausen [18]. On the other hand, the irrepresentability condition for the Lasso requires only that $2|\rho| < 1$, i.e., $\rho \in (-0.5, 0.5)$. Thus, in the regime $|\rho| \in [0.2017, 0.5)$, the Lasso irrepresentability condition holds while the log-determinant counterpart fails.

3.4.2 Illustration of irrepresentability: Star graphs

A second interesting example is the star-shaped graphical model, illustrated in Figure 2(b), which consists of a single hub node connected to the rest of the spoke nodes. We consider a four node graph, with vertex set $V = \{1, 2, 3, 4\}$ and edge-set $E = \{(1, s) \mid s \in \{2, 3, 4\}\}$. The covariance matrix Σ^* is parameterized the correlation parameter $\rho \in [-1, 1]$: the diagonal entries are set to $\Sigma^*_{ii} = 1$, for all $i \in V$; the entries corresponding to edges are set to $\Sigma^*_{ij} = \rho$ for $(i, j) \in E$; while the non-edge entries are set as $\Sigma^*_{ij} = \rho^2$ for $(i, j) \notin E$. Consequently, for this particular

example, Assumption 1 reduces to the constraint $|\rho|(|\rho|+2) < 1$, which holds for all $\rho \in (-0.414, 0.414)$. The irrepresentability condition for the Lasso on the other hand allows the full range $\rho \in (-1, 1)$. Thus there is again a regime, $|\rho| \in [0.414, 1)$, where the Lasso irrepresentability condition holds while the log-determinant counterpart fails.

3.5 Rates in Frobenius and spectral norm

We now derive some corollaries of Theorem 1 concerning estimation of Θ^* in Frobenius norm, as well as the spectral norm. Recall that $s = |E(\Theta^*)|$ denotes the total number of off-diagonal non-zeros in Θ^*.

Corollary 3. *Under the same assumptions as Theorem 1, with probability at least $1 - 1/p^{\tau-2}$, the estimator $\widehat{\Theta}$ satisfies*

$$\|\widehat{\Theta} - \Theta^*\|_F \leq \{2K_{\Gamma^*}(1 + \frac{8}{\alpha})\} \sqrt{s+p}\, \bar{\delta}_f(n, p^\tau), \qquad \text{and} \tag{41a}$$

$$\|\widehat{\Theta} - \Theta^*\|_2 \leq \{2K_{\Gamma^*}(1 + \frac{8}{\alpha})\} \min\{\sqrt{s+p},\, d\}\, \bar{\delta}_f(n, p^\tau). \tag{41b}$$

Proof. With the shorthand notation $\nu := 2K_{\Gamma^*}(1 + 8/\alpha)\ \bar{\delta}_f(n, p^\tau)$, Theorem 1 guarantees that, with probability at least $1 - 1/p^{\tau-2}$, $\|\widehat{\Theta} - \Theta^*\|_\infty \leq \nu$. Since the edge set of $\widehat{\Theta}$ is a subset of that of Θ^*, and Θ^* has at most $p + s$ non-zeros (including the diagonal), we conclude that

$$\begin{aligned} \|\widehat{\Theta} - \Theta^*\|_F &= \Big[\sum_{i=1}^{p} (\widehat{\Theta}_{ii} - \Theta^*_{ii})^2 + \sum_{(i,j)\in E} (\widehat{\Theta}_{ij} - \Theta^*_{ij})^2\Big]^{1/2} \\ &\leq \nu\sqrt{s+p}, \end{aligned}$$

from which the bound (41a) follows. On the other hand, for a symmetric matrix, we have

$$\|\widehat{\Theta} - \Theta^*\|_2 \leq \|\widehat{\Theta} - \Theta^*\|_\infty \leq d\nu, \tag{42}$$

using the definition of the ν_∞-operator norm, and the fact that $\widehat{\Theta}$ and Θ^* have at most d non-zeros per row. Since the Frobenius norm upper bounds the spectral norm, the bound (41b) follows.

□

3.5.1 Exponential-type tails

For the exponential tail function case where the rescaled random variables $X_i/\sqrt{\Sigma^*_{ii}}$ are sub-Gaussian with parameter σ, we can use the expression (32) for the inverse function $\bar{\delta}_f$ to derive rates in Frobenius and spectral norms. When the quantities $K_{\Gamma^*}, K_{\Sigma^*}, \alpha$ remain constant, these bounds can be summarized succinctly as a sample size $n = \Omega(d^2 \log p)$ is sufficient to guarantee the bounds

$$\|\widehat{\Theta} - \Theta^*\|_F = \mathcal{O}\Big(\sqrt{\frac{(s+p)\ \log p}{n}}\Big), \quad \text{and} \tag{43a}$$

$$\|\widehat{\Theta} - \Theta^*\|_2 = \mathcal{O}\Big(\sqrt{\frac{\min\{s+p,\, d^2\}\ \log p}{n}}\Big), \tag{43b}$$

with probability at least $1 - 1/p^{\tau-2}$.

3.5.2 Polynomial-type tails

Similarly, let us again consider the polynomial tail case, in which the rescaled variates $X_i/\sqrt{\Sigma^*_{ii}}$ have bounded $4m^{th}$ moments and the samples are drawn i.i.d. Using the expression (34) for the inverse function we can derive rates in

the Frobenius and spectral norms. When the quantities $K_{\Gamma^*}, K_{\Sigma^*}, \alpha$ are viewed as constant, we are guaranteed that a sample size $n = \Omega(d^2\, p^{\tau/m})$ is sufficient to guarantee the bounds

$$\|\widehat{\Theta} - \Theta^*\|_F = \mathcal{O}\left(\sqrt{\frac{(s+p)\, p^{\tau/m}}{n}}\right), \text{ and} \tag{44a}$$

$$\|\widehat{\Theta} - \Theta^*\|_2 = \mathcal{O}\left(\sqrt{\frac{\min\{s+p,\, d^2\}\, p^{\tau/m}}{n}}\right), \tag{44b}$$

with probability at least $1 - 1/p^{\tau-2}$.

3.6 Rates for the covariance matrix estimate

Finally, we describe some bounds on the estimation of the covariance matrix Σ^*. By Lemma 3, the estimated concentration matrix $\widehat{\Theta}$ is positive definite, and hence can be inverted to obtain an estimate of the covariance matrix, which we denote as $\widehat{\widehat{\Sigma}} := (\widehat{\Theta})^{-1}$.

Corollary 4. *Under the same assumptions as Theorem 1, with probability at least $1 - 1/p^{\tau-2}$, the following bounds hold.*

(a) The element-wise ℓ_∞ norm of the deviation $(\widehat{\widehat{\Sigma}} - \Sigma^*)$ satisfies the bound

$$\|\widehat{\widehat{\Sigma}} - \Sigma^*\|_\infty \leq C_3, [\bar{\delta}_f(n, p^\tau)] + C_4 d\, [\bar{\delta}_f(n, p^\tau)]^2 \tag{45}$$

where $C_3 = 2K_{\Sigma^*}^2 K_{\Gamma^*}\left(1 + \frac{8}{\alpha}\right)$ and $C_4 = 6K_{\Sigma^*}^3 K_{\Gamma^*}^2\left(1 + \frac{8}{\alpha}\right)^2$.

(b) The ℓ_2 operator-norm of the deviation $(\widehat{\widehat{\Sigma}} - \Sigma^*)$ satisfies the bound

$$\|\widehat{\widehat{\Sigma}} - \Sigma^*\|_2 \leq C_3\, d\, [\bar{\delta}_f(n, p^\tau)] + C_4 d^2\, [\bar{\delta}_f(n, p^\tau)]^2. \tag{46}$$

The proof involves certain lemmata and derivations that are parts of the proofs of Theorems 1 and 2, so that we defer it to Section 4.5.

4 Proofs of main result

In this section, we work through the proofs of Theorems 1 and 2. We break down the proofs into a sequence of lemmas, with some of the more technical aspects deferred to appendices.

Our proofs are based on a technique that we call a *primal-dual witness method*, used previously in analysis of the Lasso [23]. It involves following a specific sequence of steps to construct a pair $(\widetilde{\Theta}, \widetilde{Z})$ of symmetric matrices that together satisfy the optimality conditions associated with the convex program (11) *with high probability*. Thus, when the constructive procedure succeeds, $\widetilde{\Theta}$ is *equal* to the unique solution $\widehat{\Theta}$ of the convex program (11), and $\widetilde{Z}$ is an optimal solution to its dual. In this way, the estimator $\widehat{\Theta}$ inherits from $\widetilde{\Theta}$ various optimality properties in terms of its distance to the truth Θ^*, and its recovery of the signed sparsity pattern. To be clear, our procedure for constructing $\widetilde{\Theta}$ is *not* a practical algorithm for solving the log-determinant problem (11), but rather is used as a proof technique for certifying the behavior of the M-estimator (11).

4.1 Primal-dual witness approach

As outlined above, at the core of the primal-dual witness method are the standard convex optimality conditions that characterize the optimum $\widehat{\Theta}$ of the convex program (11). For future reference, we note that the sub-differential of the

norm $\|\cdot\|_{1,\text{off}}$ evaluated at some Θ consists the set of all symmetric matrices $Z \in \mathbb{R}^{p\times p}$ such that

$$Z_{ij} \;=\; \begin{cases} 0 & \text{if } i = j \\ \text{sign}(\Theta_{ij}) & \text{if } i \neq j \text{ and } \Theta_{ij} \neq 0 \\ \in [-1,+1] & \text{if } i \neq j \text{ and } \Theta_{ij} = 0. \end{cases} \tag{47}$$

The following result is proved in Appendix A:

Lemma 3. *For any $\lambda_n > 0$ and sample covariance $\widehat{\Sigma}$ with strictly positive diagonal, the ℓ_1-regularized log-determinant problem (11) has a unique solution $\widehat{\Theta} \succ 0$ characterized by*

$$\widehat{\Sigma} - \widehat{\Theta}^{-1} + \lambda_n \widehat{Z} \;=\; 0, \tag{48}$$

where $\widehat{Z}$ is an element of the subdifferential $\partial\|\widehat{\Theta}\|_{1,\text{off}}$.

Based on this lemma, we construct the primal-dual witness solution $(\widetilde{\Theta}, \widetilde{Z})$ as follows:

(a) We determine the matrix $\widetilde{\Theta}$ by solving the restricted log-determinant problem

$$\widetilde{\Theta} \;:=\; \arg\min_{\Theta\succ 0,\ \Theta_{S^c}=0} \big\{ \langle\!\langle \Theta,\ \widehat{\Sigma} \rangle\!\rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \big\}. \tag{49}$$

Note that by construction, we have $\widetilde{\Theta} \succ 0$, and moreover $\widetilde{\Theta}_{S^c} = 0$.

(b) We choose $\widetilde{Z}_S$ as a member of the sub-differential of the regularizer $\|\cdot\|_{1,\text{off}}$, evaluated at $\widetilde{\Theta}$.

(c) We set $\widetilde{Z}_{S^c}$ as

$$\widetilde{Z}_{S^c} \;=\; \frac{1}{\lambda_n}\big\{ -\widehat{\Sigma}_{S^c} + [\widetilde{\Theta}^{-1}]_{S^c} \big\}, \tag{50}$$

which ensures that constructed matrices $(\widetilde{\Theta}, \widetilde{Z})$ satisfy the optimality condition (48).

(d) We verify the *strict dual feasibility* condition

$$|\widetilde{Z}_{ij}| \;<\; 1 \quad \text{for all } (i,j) \in S^c.$$

To clarify the nature of the construction, steps (a) through (c) suffice to obtain a pair $(\widetilde{\Theta}, \widetilde{Z})$ that satisfy the optimality conditions (48), but do *not* guarantee that $\widetilde{Z}$ is an element of sub-differential $\partial\|\widetilde{\Theta}\|_{1,\text{off}}$. By construction, specifically step (b) of the construction ensures that the entries $\widetilde{Z}$ in S satisfy the sub-differential conditions, since $\widetilde{Z}_S$ is a member of the sub-differential of $\partial\|\widetilde{\Theta}_S\|_{1,\text{off}}$. The purpose of step (d), then, is to verify that the remaining elements of $\widetilde{Z}$ satisfy the necessary conditions to belong to the sub-differential.

If the primal-dual witness construction succeeds, then it acts as a *witness* to the fact that the solution $\widetilde{\Theta}$ to the restricted problem (49) is equivalent to the solution $\widehat{\Theta}$ to the original (unrestricted) problem (11). We exploit this fact in our proofs of Theorems 1 and 2 that build on this: we first show that the primal-dual witness technique succeeds with high-probability, from which we can conclude that the support of the optimal solution $\widehat{\Theta}$ is contained within the support of the true Θ^*. In addition, we exploit the characterization of $\widehat{\Theta}$ provided by the primal-dual witness construction to establish the elementwise ℓ_∞ bounds claimed in Theorem 1. Theorem 2 requires checking, in addition, that certain sign consistency conditions hold, for which we require lower bounds on the value of the minimum value $\theta_{\min}$.

In the analysis to follow, some additional notation is useful. We let W denote the "effective noise" in the sample covariance matrix $\widehat{\Sigma}$, namely

$$W \;:=\; \widehat{\Sigma} - (\Theta^*)^{-1}. \tag{51}$$

Second, we use $\Delta = \widetilde{\Theta} - \Theta^*$ to measure the discrepancy between the primal witness matrix $\widetilde{\Theta}$ and the truth Θ^*. Finally, recall the log-determinant barrier g from equation (6). We let $R(\Delta)$ denote the difference of the gradient $\nabla g(\widetilde{\Theta}) = \widetilde{\Theta}^{-1}$ from its first-order Taylor expansion around Θ^*. Using known results on the first and second derivatives of the log-determinant function (see p. 641 in Boyd and Vandenberghe [3]), this remainder takes the form

$$R(\Delta) \;=\; \widetilde{\Theta}^{-1} - {\Theta^*}^{-1} + {\Theta^*}^{-1}\Delta{\Theta^*}^{-1}. \tag{52}$$

4.2 Auxiliary results

We begin with some auxiliary lemmata, required in the proofs of our main theorems. In Section 4.2.1, we provide sufficient conditions on the quantities W and R for the strict dual feasibility condition to hold. In Section 4.2.2, we control the remainder term $R(\Delta)$ in terms of Δ, while in Section 4.2.3, we control Δ itself, providing elementwise ℓ_∞ bounds on Δ. In Section 4.2.4, we show that under appropriate conditions on the minimum value $\theta_{\min}$, the bounds in the earlier lemmas guarantee that the sign consistency condition holds. All of the analysis in these sections is *deterministic* in nature. In Section 4.2.5, we turn to the probabilistic component of the analysis, providing control of the noise W in the sample covariance matrix. Finally, the proofs of Theorems 1 and 2 follows by using this probabilistic control of W and the stated conditions on the sample size to show that the deterministic conditions hold with high probability.

4.2.1 Sufficient conditions for strict dual feasibility

We begin by stating and proving a lemma that provides sufficient (deterministic) conditions for strict dual feasibility to hold, so that $\|\widetilde{Z}_{S^c}\|_\infty < 1$.

Lemma 4 (Strict dual feasibility)**.** *Suppose that*

$$\max\left\{\|W\|_\infty,\ \|R(\Delta)\|_\infty\right\} \quad \leq \quad \frac{\alpha\,\lambda_n}{8}. \tag{53}$$

Then the matrix $\widetilde{Z}_{S^c}$ constructed in step (c) satisfies $\|\widetilde{Z}_{S^c}\|_\infty < 1$, and therefore $\widetilde{\Theta} = \widehat{\Theta}$.

Proof. Using the definitions (51) and (52), we can re-write the stationary condition (48) in an alternative but equivalent form

$$\Theta^{*-1}\Delta\Theta^{*-1} + W - R(\Delta) + \lambda_n\widetilde{Z} \quad = 0. \tag{54}$$

This is a linear-matrix equality, which can be re-written as an ordinary linear equation by "vectorizing" the matrices. We use the notation $\mathrm{vec}(A)$, or equivalently $\overline{A}$ for the p^2-vector version of the matrix $A \in \mathbb{R}^{p\times p}$, obtained by stacking up the rows into a single column vector. In vectorized form, we have

$$\mathrm{vec}\left(\Theta^{*-1}\Delta\Theta^{*-1}\right) = \left(\Theta^{*-1}\otimes\Theta^{*-1}\right)\overline{\Delta} \ = \ \Gamma^*\overline{\Delta}.$$

In terms of the disjoint decomposition S and S^c, equation (54) can be re-written as two blocks of linear equations as follows:

$$\Gamma^*_{SS}\overline{\Delta}_S + \overline{W}_S - \overline{R}_S + \lambda_n\overline{\widetilde{Z}}_S \quad = \quad 0 \tag{55a}$$

$$\Gamma^*_{S^cS}\overline{\Delta}_S + \overline{W}_{S^c} - \overline{R}_{S^c} + \lambda_n\overline{\widetilde{Z}}_{S^c} \quad = \quad 0. \tag{55b}$$

Here we have used the fact that $\Delta_{S^c} = 0$ by construction.

Since Γ^*_{SS} is invertible, we can solve for $\overline{\Delta}_S$ from equation (55a) as follows:

$$\overline{\Delta}_S \quad = \quad \left(\Gamma^*_{SS}\right)^{-1}\left[-\overline{W}_S + \overline{R}_S - \lambda_n\overline{\widetilde{Z}_S}\right].$$

Substituting this expression into equation (55b), we can solve for $\widetilde{Z}_{S^c}$ as follows:

$$\begin{aligned}\overline{\widetilde{Z}}_{S^c} \quad &= \quad -\frac{1}{\lambda_n}\Gamma^*_{S^cS}\overline{\Delta}_S + \frac{1}{\lambda_n}\overline{R}_{S^c} - \frac{1}{\lambda_n}\overline{W}_{S^c} \\ &= \quad -\frac{1}{\lambda_n}\Gamma^*_{S^cS}\left(\Gamma^*_{SS}\right)^{-1}(\overline{W}_S - \overline{R}_S) + \Gamma^*_{S^cS}\left(\Gamma^*_{SS}\right)^{-1}\overline{\widetilde{Z}}_S - \frac{1}{\lambda_n}(\overline{W}_{S^c} - \overline{R}_{S^c}).\end{aligned} \tag{56}$$

Taking the ℓ_∞ norm of both sides yields

$$\|\bar{\widetilde{Z}}_{S^c}\|_\infty \leq \frac{1}{\lambda_n}\|\Gamma^*_{S^cS}\big(\Gamma^*_{SS}\big)^{-1}\|_\infty(\|\overline{W}_S\|_\infty + \|\overline{R}_S\|_\infty) + \|\Gamma^*_{S^cS}\big(\Gamma^*_{SS}\big)^{-1}\|_\infty\|\bar{\widetilde{Z}}_S\|_\infty + \frac{1}{\lambda_n}(\|\overline{W}_S\|_\infty + \|\overline{R}_S\|_\infty).$$

Recalling Assumption 1—namely, that $\|\Gamma^*_{S^cS}\big(\Gamma^*_{SS}\big)^{-1}\|_\infty \leq (1-\alpha)$—we have

$$\|\bar{\widetilde{Z}}_{S^c}\|_\infty \quad \leq \quad \frac{2-\alpha}{\lambda_n}\left(\|\overline{W}_S\|_\infty + \|\overline{R}_S\|_\infty\right) + (1-\alpha),$$

where we have used the fact that $\|\bar{\widetilde{Z}}_S\|_\infty \leq 1$, since $\widetilde{Z}$ belongs to the sub-differential of the norm $\|\cdot\|_{1,\text{off}}$ by construction. Finally, applying assumption (53) from the lemma statement, we have

$$\begin{aligned}\|\bar{\widetilde{Z}}_{S^c}\|_\infty &\leq \frac{(2-\alpha)}{\lambda_n}\left(\frac{\alpha\lambda_n}{4}\right) + (1-\alpha) \\ &\leq \frac{\alpha}{2} + (1-\alpha) \;<\; 1,\end{aligned}$$

as claimed.

□

4.2.2 Control of remainder term

Our next step is to relate the behavior of the remainder term (52) to the deviation $\Delta = \widetilde{\Theta} - \Theta^*$.

Lemma 5 (Control of remainder). *Suppose that the elementwise ℓ_∞ bound $\|\Delta\|_\infty \leq \frac{1}{3\,K_{\Sigma^*}d}$ holds. Then:*

$$R(\Delta) \quad = \quad \Theta^{*-1}\Delta\Theta^{*-1}\Delta J\Theta^{*-1}, \tag{57}$$

where $J := \sum_{k=0}^{\infty}(-1)^k\left(\Theta^{*-1}\Delta\right)^k$ has norm $\|J^T\|_\infty \leq 3/2$. Moreover, in terms of the elementwise ℓ_∞-norm, we have

$$\|R(\Delta)\|_\infty \quad \leq \quad \frac{3}{2}d\|\Delta\|^2_\infty\, K^3_{\Sigma^*}. \tag{58}$$

We provide the proof of this lemma in Appendix B. It is straightforward, based on standard matrix expansion techniques.

4.2.3 Sufficient conditions for ℓ_∞ bounds

Our next lemma provides control on the deviation $\Delta = \widetilde{\Theta} - \Theta^*$, measured in elementwise ℓ_∞ norm.

Lemma 6 (Control of Δ). *Suppose that*

$$r := 2K_{\Gamma^*}\left(\|W\|_\infty + \lambda_n\right) \quad \leq \quad \min\left\{\frac{1}{3K_{\Sigma^*}d},\ \frac{1}{3K^3_{\Sigma^*}\,K_{\Gamma^*}d}\right\}. \tag{59}$$

Then we have the elementwise ℓ_∞ bound

$$\|\Delta\|_\infty = \|\widetilde{\Theta} - \Theta^*\|_\infty \quad \leq \quad r. \tag{60}$$

We prove the lemma in Appendix C; at a high level, the main steps involved are the following. We begin by noting that $\widetilde{\Theta}_{S^c} = \Theta^*_{S^c} = 0$, so that $\|\Delta\|_\infty = \|\Delta_S\|_\infty$. Next, we characterize $\widetilde{\Theta}_S$ in terms of the zero-gradient condition associated with the restricted problem (49). We then define a continuous map $F : \Delta_S \mapsto F(\Delta_S)$ such that its fixed points are equivalent to zeros of this gradient expression in terms of $\Delta_S = \widetilde{\Theta}_S - \Theta^*_S$. We then show that the function F maps the ℓ_∞-ball

$$\mathbb{B}(r) \quad := \quad \{\Theta_S \mid \|\Theta_S\|_\infty \leq r\}, \qquad \text{with } r := 2K_{\Gamma^*}\left(\|W\|_\infty + \lambda_n\right), \tag{61}$$

onto itself. Finally, with these results in place, we can apply Brouwer's fixed point theorem (e.g., p. 161; Ortega and Rheinboldt [20]) to conclude that F does indeed have a fixed point inside $\mathbb{B}(r)$.

4.2.4 Sufficient conditions for sign consistency

We now show how a lower bound on the minimum value $\theta_{\min}$, when combined with Lemma 6, allows us to guarantee *sign consistency* of the primal witness matrix $\widetilde{\Theta}_S$.

Lemma 7 (Sign Consistency)**.** *Suppose the minimum absolute value $\theta_{\min}$ of non-zero entries in the true concentration matrix Θ^* is lower bounded as*

$$\theta_{\min} \quad \geq \quad 4K_{\Gamma^*}\big(\|W\|_\infty + \lambda_n\big), \tag{62}$$

then sign$(\widetilde{\Theta}_S)$ = sign(Θ^*_S) holds.

This claim follows from the bound (62) combined with the bound (60) ,which together imply that for all $(i, j) \in S$, the estimate $\widetilde{\Theta}_{ij}$ cannot differ enough from Θ^*_{ij} to change sign.

4.2.5 Control of noise term

The final ingredient required for the proofs of Theorems 1 and 2 is control on the sampling noise $W = \widehat{\Sigma} - \Sigma^*$. This control is specified in terms of the decay function f from equation (12).

Lemma 8 (Control of Sampling Noise)**.** *For any $\tau > 2$ and sample size n such that $\bar{\delta}_f(n, p^\tau) \leq 1/v_*$, we have*

$$\mathbb{P}\Big[\|W\|_\infty \geq \bar{\delta}_f(n, p^\tau)\Big] \quad \leq \quad \frac{1}{p^{\tau-2}} \;\to\; 0. \tag{63}$$

Proof. Using the definition (12) of the decay function f, and applying the union bound over all p^2 entries of the noise matrix, we obtain that for all $\delta \leq 1/v_*$,

$$\mathbb{P}\big[\max_{i,j} |W_{ij}| \geq \delta\big] \quad \leq \quad p^2/f(n, \delta).$$

Setting $\delta = \bar{\delta}_f(n, p^\tau)$ yields that

$$\mathbb{P}\big[\max_{i,j} |W_{ij}| \geq \bar{\delta}_f(n, p^\tau)\big] \quad \leq \quad p^2/\big[f(n, \bar{\delta}_f(n, p^\tau))\big] \;=\; 1/p^{\tau-2},$$

as claimed. Here the last equality follows since $f(n, \bar{\delta}_f(n, p^\tau)) = p^\tau$, using the definition (14) of the inverse function $\bar{\delta}_f$. □

4.3 Proof of Theorem 1

We now have the necessary ingredients to prove Theorem 1. We first show that with high probability the witness matrix $\widetilde{\Theta}$ is equal to the solution $\widehat{\Theta}$ to the original log-determinant problem (11), in particular by showing that the primal-dual witness construction (described in in Section 4.1) succeeds with high probability. Let $\mathcal{A}$ denote the event that $\|W\|_\infty \leq \bar{\delta}_f(n, p^\tau)$. Using the monotonicity of the inverse tail function (15), the lower lower bound (29) on the sample size n implies that $\bar{\delta}_f(n, p^\tau) \leq 1/v_*$. Consequently, Lemma 8 implies that $\mathbb{P}(\mathcal{A}) \geq 1 - \frac{1}{p^{\tau-2}}$. Accordingly, we condition on the event $\mathcal{A}$ in the analysis to follow.

We proceed by verifying that assumption (53) of Lemma 4 holds. Recalling the choice of regularization penalty $\lambda_n = (8/\alpha)\,\bar{\delta}_f(n, p^\tau)$, we have $\|W\|_\infty \leq (\alpha/8)\lambda_n$. In order to establish condition (53) it remains to establish the bound $\|R(\Delta)\|_\infty \leq \frac{\alpha\,\lambda_n}{8}$. We do so in two steps, by using Lemmas 6 and 5 consecutively. First, we show that the precondition (59) required for Lemma 6 to hold is satisfied under the specified conditions on n and λ_n. From Lemma 8 and our choice of regularization constant $\lambda_n = (8/\alpha)\,\bar{\delta}_f(n, p^\tau)$,

$$2K_{\Gamma^*}\big(\,\|W\|_\infty + \lambda_n\big) \quad \leq \quad 2K_{\Gamma^*}\Big(1 + \frac{8}{\alpha}\Big)\,\bar{\delta}_f(n, p^\tau),$$

provided $\bar{\delta}_f(n, p^\tau) \leq 1/v_*$. From the lower bound (29) and the monotonicity (15) of the tail inverse functions, we have

$$2K_{\Gamma^*}\Big(1+\frac{8}{\alpha}\Big)\,\bar{\delta}_f(n,p^\tau) \;\;\leq\;\; \min\Big\{\frac{1}{3K_{\Sigma^*}d},\;\frac{1}{3K_{\Sigma^*}^3\;K_{\Gamma^*}d}\Big\}, \tag{64}$$

showing that the assumptions of Lemma 6 are satisfied. Applying this lemma, we conclude that

$$\|\Delta\|_\infty \;\;\leq\;\; 2K_{\Gamma^*}\big(\,\|W\|_\infty+\lambda_n\big) \;\leq\; 2K_{\Gamma^*}\Big(1+\frac{8}{\alpha}\Big)\,\bar{\delta}_f(n,p^\tau). \tag{65}$$

Turning next to Lemma 5, we see that its assumption $\|\Delta\|_\infty \leq \frac{1}{3\,K_{\Sigma^*}d}$ holds, by applying equations (64) and (65). Consequently, we have

$$\begin{aligned}
\|R(\Delta)\|_\infty &\leq \frac{3}{2}\,d\,\|\Delta\|_\infty^2\,K_{\Sigma^*}^3 \\
&\leq 6K_{\Sigma^*}^3K_{\Gamma^*}^2\,d\Big(1+\frac{8}{\alpha}\Big)^2[\bar{\delta}_f(n,p^\tau)]^2 \\
&= \Big\{6K_{\Sigma^*}^3K_{\Gamma^*}^2\,d\Big(1+\frac{8}{\alpha}\Big)^2\bar{\delta}_f(n,p^\tau)\Big\}\frac{\alpha\lambda_n}{8} \\
&\leq \frac{\alpha\lambda_n}{8},
\end{aligned}$$

as required, where the final inequality follows from our condition (29) on the sample size, and the monotonicity property (15).

Overall, we have shown that the assumption (53) of Lemma 4 holds, allowing us to conclude that $\widetilde{\Theta} = \widehat{\Theta}$. The estimator $\widehat{\Theta}$ then satisfies the ℓ_∞-bound (65) of $\widetilde{\Theta}$, as claimed in Theorem 1(a), and moreover, we have $\widehat{\Theta}_{S^c} = \widetilde{\Theta}_{S^c} = 0$, as claimed in Theorem 1(b). Since the above was conditioned on the event $\mathcal{A}$, these statements hold with probability $\mathbb{P}(\mathcal{A}) \geq 1 - \frac{1}{p^{\tau-2}}$.

4.4 Proof of Theorem 2

We now turn to the proof of Theorem 2. A little calculation shows that the assumed lower bound (37) on the sample size n and the monotonicity property (15) together guarantee that

$$\theta_{\min} \;\;>\;\; 4K_{\Gamma^*}\Big(1+\frac{8}{\alpha}\Big)\,\bar{\delta}_f(n,p^\tau)$$

Proceeding as in the proof of Theorem 1, with probability at least $1 - 1/p^{\tau-2}$, we have the equality $\widetilde{\Theta} = \widehat{\Theta}$, and also that $\|\widetilde{\Theta} - \Theta^*\|_\infty \leq \theta_{\min}/2$. Consequently, Lemma 7 can be applied, guaranteeing that $\operatorname{sign}(\Theta^*_{ij}) = \operatorname{sign}(\widetilde{\Theta}_{ij})$ for all $(i,j) \in E$. Overall, we conclude that with probability at least $1 - 1/p^{\tau-2}$, the sign consistency condition $\operatorname{sign}(\Theta^*_{ij}) = \operatorname{sign}(\widehat{\Theta}_{ij})$ holds for all $(i,j) \in E$, as claimed.

4.5 Proof of Corollary 4

With the shorthand $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$, we have

$$\widehat{\widehat{\Sigma}} - \Sigma^* = (\Theta^* + \widehat{\Delta})^{-1} - \big(\Theta^*\big)^{-1}.$$

From the definition (52) of the residual $R(\cdot)$, this difference can be written as

$$\widehat{\widehat{\Sigma}} - \Sigma^* \;\;=\;\; -\Theta^{*-1}\widehat{\Delta}\Theta^{*-1} + R(\widehat{\Delta}). \tag{66}$$

Proceeding as in the proof of Theorem 1 we condition on the event $\mathcal{A} = \{\|W\|_\infty \leq \bar{\delta}_f(n, p^\tau)\}$, and which holds with probability $\mathbb{P}(\mathcal{A}) \geq 1 - \frac{1}{p^{\tau-2}}$. As in the proof of that theorem, we are guaranteed that the assumptions of Lemma 5 are satisfied, allowing us to conclude

$$R(\widehat{\Delta}) = \Theta^{*-1}\widehat{\Delta}\Theta^{*-1}\widehat{\Delta}J\Theta^{*-1}, \tag{67}$$

where $J := \sum_{k=0}^{\infty}(-1)^k\left(\Theta^{*-1}\widehat{\Delta}\right)^k$ has norm $\|J^T\|_\infty \leq 3/2$.

We begin by proving the bound (45). From equation (66), we have $\|\widehat{\widehat{\Sigma}} - \Sigma^*\|_\infty \leq \|L(\widehat{\Delta})\|_\infty + \|R(\widehat{\Delta})\|_\infty$. From Lemma 5, we have the elementwise ℓ_∞-norm bound

$$\|R(\widehat{\Delta})\|_\infty \quad \leq \quad \frac{3}{2}d\|\widehat{\Delta}\|_\infty^2 \, K_{\Sigma^*}^3.$$

The quantity $L(\widehat{\Delta})$ in turn can be bounded as follows,

$$\begin{aligned}
\|L(\widehat{\Delta})\|_\infty &= \max_{ij}\left|e_i^T\Theta^{*-1}\widehat{\Delta}\Theta^{*-1}e_j\right| \\
&\leq \max_i \|e_i^T\Theta^{*-1}\|_1 \max_j \|\widehat{\Delta}\Theta^{*-1}e_j\|_\infty \\
&\leq \max_i \|e_i^T\Theta^{*-1}\|_1 \|\widehat{\Delta}\|_\infty \| \max_j \|\Theta^{*-1}e_j\|_1
\end{aligned}$$

where we used the inequality that $\|\widehat{\Delta}u\|_\infty \leq \|\widehat{\Delta}\|_\infty\|u\|_1$. Simplifying further, we obtain

$$\begin{aligned}
\|L(\widehat{\Delta})\|_\infty &\leq \|\Theta^{*-1}\|_\infty\|\widehat{\Delta}\|_\infty\|\Theta^{*-1}\|_1 \\
&\leq \|\Theta^{*-1}\|_\infty^2\|\widehat{\Delta}\|_\infty \\
&\leq K_{\Sigma^*}^2\|\widehat{\Delta}\|_\infty,
\end{aligned}$$

where we have used the fact that $\|\Theta^{*-1}\|_1 = \|[\Theta^{*-1}]^T\|_\infty = \|\Theta^{*-1}\|_\infty$, which follows from the symmetry of Θ^{*-1}. Combining the pieces, we obtain

$$\begin{aligned}
\|\widehat{\widehat{\Sigma}} - \Sigma^*\|_\infty &\leq \|L(\widehat{\Delta})\|_\infty + \|R(\widehat{\Delta})\|_\infty \\
&\leq K_{\Sigma^*}^2\|\widehat{\Delta}\|_\infty + \frac{3}{2}dK_{\Sigma^*}^3\|\widehat{\Delta}\|_\infty^2.
\end{aligned} \tag{68}$$

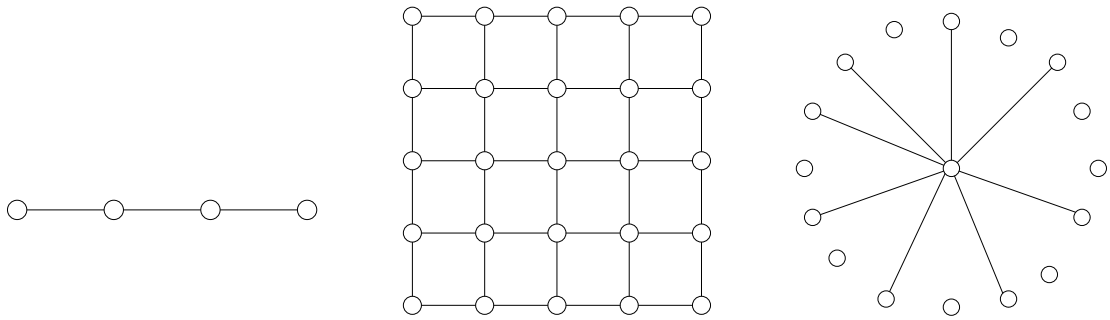
The claim then follows from the elementwise ℓ_∞-norm bound (30) from Theorem 1.

Next, we establish the bound (46) in spectral norm. Taking the ℓ_∞ operator norm of both sides of equation (66) yields the inequality $\|\widehat{\widehat{\Sigma}} - \Sigma^*\|_\infty \leq \|L(\widehat{\Delta})\|_\infty + \|R(\widehat{\Delta})\|_\infty$. Using the expansion (67), and the sub-multiplicativity of the ℓ_∞ operator norm, we obtain

$$\begin{aligned}
\|R(\widehat{\Delta})\|_\infty &\leq \|\Theta^{*-1}\|_\infty\|\widehat{\Delta}\|_\infty\|\Theta^{*-1}\|_\infty\|\widehat{\Delta}\|_\infty\|J\|_\infty\|\Theta^{*-1}\|_\infty \\
&\leq \|\Theta^{*-1}\|_\infty^3\|J\|_\infty\|\widehat{\Delta}\|_\infty^2 \\
&\leq \frac{3}{2}K_{\Sigma^*}^3\|\widehat{\Delta}\|_\infty^2,
\end{aligned}$$

where the last inequality uses the bound $\|J\|_\infty \leq 3/2$. (Proceeding as in the proof of Lemma 5, this bound holds conditioned on $\mathcal{A}$, and for the sample size specified in the theorem statement.) In turn, the term $L(\widehat{\Delta})$ can be bounded as

$$\begin{aligned}
\|L(\Delta)\|_\infty &\leq \|\Theta^{*-1}\widehat{\Delta}\Theta^{*-1}\|_\infty \\
&\leq \|\Theta^{*-1}\|_\infty^2\|\widehat{\Delta}\|_\infty \\
&\leq K_{\Sigma^*}^2\|\widehat{\Delta}\|_\infty,
\end{aligned}$$



(a) (b) (c)

Figure 3. Illustrations of different graph classes used in simulations. (a) Chain ($d = 2$). (b) Four-nearest neighbor grid ($d = 4$) and (c) Star-shaped graph ($d \in \{1, \ldots, p-1\}$).

where the second inequality uses the sub-multiplicativity of the ℓ_∞-operator norm. Combining the pieces yields

$$|||\widehat{\widehat{\Sigma}} - \Sigma^*|||_\infty \quad \leq \quad |||L(\widehat{\Delta})|||_\infty + |||R(\widehat{\Delta})|||_\infty \quad \leq \quad K_{\Sigma^*}^2 |||\widehat{\Delta}|||_\infty + \frac{3}{2} K_{\Sigma^*}^3 |||\widehat{\Delta}|||_\infty^2. \tag{69}$$

Conditioned on the event $\mathcal{A}$, we have the bound (42) on the ℓ_∞-operator norm

$$|||\widehat{\Delta}|||_\infty \quad \leq \quad 2K_{\Gamma^*}\left(1 + \frac{8}{\alpha}\right) d\, \bar{\delta}_f(n, p^\tau).$$

Substituting this bound, as well as the elementwise ℓ_∞-norm bound (30) from Theorem 1, into the bound (69) yields the stated claim.

5 Experiments

In this section, we illustrate our results with various experimental simulations, reporting results in terms of the probability of correct model selection (Theorem 2) or the ℓ_∞-error (Theorem 1). For these illustrations, we study the case of Gaussian graphical models, and results for three different classes of graphs, namely chains, grids, and star-shaped graphs. We also consider various scalings of the quantities which affect the performance of the estimator: in addition the triple (n, p, d), we also report some results concerning the role of the parameters K_{Σ^*}, K_{Γ^*} and $\theta_{\min}$ that we have identified in the main theorems. For all results reported here, we solved the resulting ℓ_1-penalized log-determinant program (11) using the `glasso` program of Friedman et al. [10], which builds on the block co-ordinate descent algorithm of d'Asprémont et al. [8].

Figure 3 illustrates the three types of graphs used in our simulations: chain graphs (panel (a)), four-nearest neighbor lattices or grids (panel (b)), and star-shaped graphs (panel (c)). For the chain and grid graphs, the maximal node degree d is fixed by definition, to $d = 2$ for chains, and $d = 4$ for the grids. Consequently, these graphs can capture the dependence of the required sample size n only as a function of the graph size p, and the parameters $(K_{\Sigma^*}, K_{\Gamma^*}, \theta_{\min})$. The star graph allows us to vary both d and p, since the degree of the central hub can be varied between 1 and $p-1$. For each graph type, we varied the size of the graph p in different ranges, from $p = 64$ upwards to $p = 375$.

For the chain and star graphs, we define a covariance matrix Σ^* with entries $\Sigma^*_{ii} = 1$ for all $i = 1, \ldots, p$, and $\Sigma^*_{ij} = \rho$ for all $(i, j) \in E$ for specific values of ρ specified below. Note that these covariance matrices are sufficient to specify the full model. For the four-nearest neighbor grid graph, we set the entries of the concentration matrix $\Theta^*_{ij} = \omega$ for $(i, j) \in E$, with the value ω specified below. In all cases, we set the regularization parameter λ_n proportional to $\sqrt{\log(p)/n}$, as suggested by Theorems 1 and 2, which is reasonable since the main purpose of these simulations is to illustrate our theoretical results. However, for general data sets, the relevant theoretical parameters cannot be computed (since the true model is unknown), so that a data-driven approach such as cross-validation might be required for selecting the regularization parameter λ_n.

Given a Gaussian graphical model instance, and the number of samples n, we drew $N = 100$ batches of n independent samples from the associated multivariate Gaussian distribution. We estimated the probability of correct model selection as the fraction of the $N = 100$ trials in which the estimator recovers the signed-edge set exactly.

Note that any multivariate Gaussian random vector is sub-Gaussian; in particular, the rescaled variates $X_i/\sqrt{\Sigma^*_{ii}}$ are sub-Gaussian with parameter $\sigma = 1$, so that the elementwise ℓ_∞-bound from Corollary 1 applies. Suppose we collect relevant parameters such as $\theta_{\min}$ and the covariance and Hessian-related terms K_{Σ^*}, K_{Γ^*} and α into a single "model-complexity" term K defined as

$$K \quad := \quad \left[(1 + 8\alpha^{-1})(\max_i \Sigma^*_{ii}) \max\{K_{\Sigma^*} K_{\Gamma^*}, K^3_{\Sigma^*} K^2_{\Gamma^*}, \frac{K_{\Gamma^*}}{d\,\theta_{\min}}\} \right]. \tag{70}$$

Then, as a corollary of Theorem 2, a sample size of order

$$n \quad = \quad \Omega\left(K^2\, d^2\, \tau \log p\right), \tag{71}$$

is sufficient for model selection consistency with probability greater than $1 - 1/p^{\tau-2}$. In the subsections to follow, we investigate how the empirical sample size n required for model selection consistency scales in terms of graph size p, maximum degree d, as well as the "model-complexity" term K defined above.

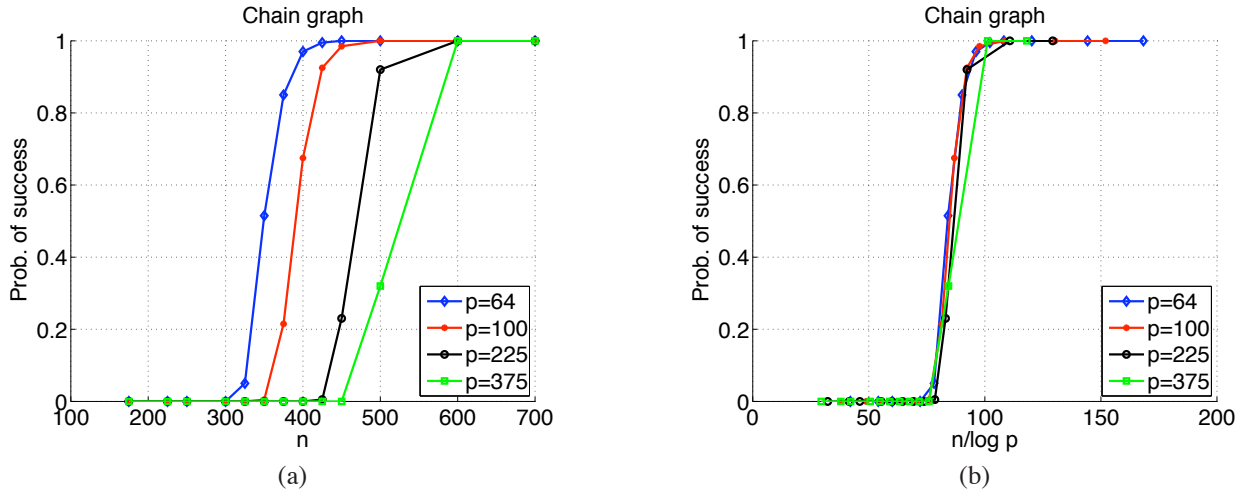


Figure 4. Simulations for chain graphs with varying number of nodes p, edge covariances $\Sigma^*_{ij} = 0.10$. Plots of probability of correct signed edge-set recovery plotted versus the ordinary sample size n in panel (a), and versus the rescaled sample size $n/\log p$ in panel (b). Each point corresponds to the average over 100 trials.

5.1 Dependence on graph size

Panel (a) of Figure 4 plots the probability of correct signed edge-set recovery against the sample size n for a chain-structured graph of three different sizes. For these chain graphs, regardless of the number of nodes p, the maximum node degree is constant $d = 2$, while the edge covariances are set as $\Sigma_{ij} = 0.2$ for all $(i, j) \in E$, so that the quantities $(K_{\Sigma^*}, K_{\Gamma^*}, \alpha)$ remain constant. Each of the curve in panel (a) corresponds to a different graph size p. For each curve, the probability of success starts at zero (for small sample sizes n), but then transitions to one as the sample size is increased. As would be expected, it is more difficult to perform model selection for larger graph sizes, so that (for instance) the curve for $p = 375$ is shifted to the right relative to the curve for $p = 64$. Panel (b) of Figure 4 replots the same data, with the horizontal axis rescaled by $(1/\log p)$. This scaling was chosen because for sub-Gaussian tails, our theory predicts that the sample size should scale logarithmically with p (see equation (71)). Consistent with this prediction, when plotted against the rescaled sample size $n/\log p$, the curves in panel (b) all stack up. Consequently,

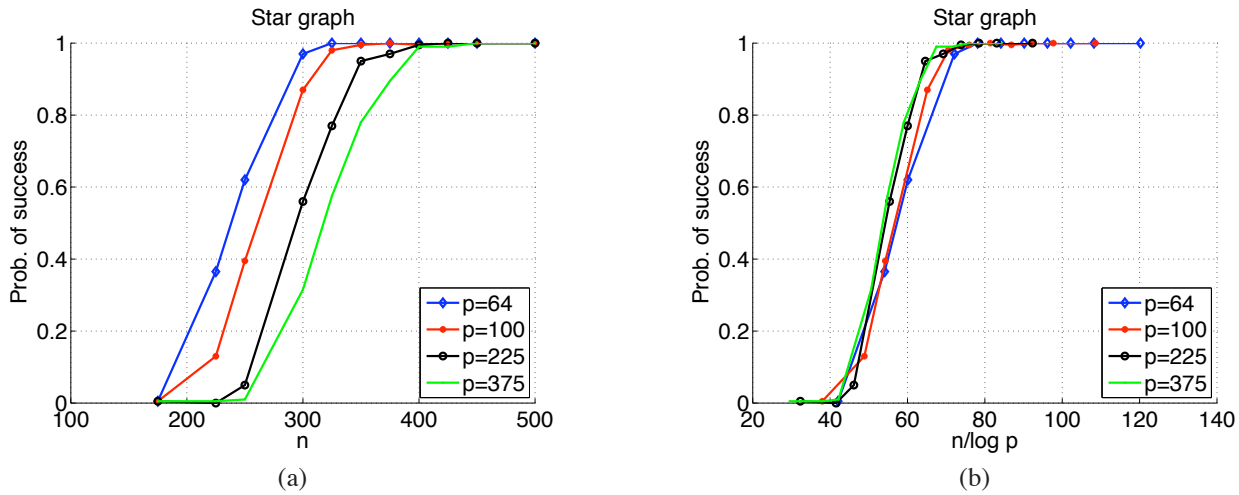


Figure 5. Simulations for a star graph with varying number of nodes p, fixed maximal degree $d = 40$, and edge covariances $\Sigma^*_{ij} = 1/16$ for all edges. Plots of probability of correct signed edge-set recovery versus the sample size n in panel (a), and versus the rescaled sample size $n/\log p$ in panel (b). Each point corresponds to the average over $N = 100$ trials.

the ratio $(n/\log p)$ acts as an effective sample size in controlling the success of model selection, consistent with the predictions of Theorem 2 for sub-Gaussian variables.

Figure 5 shows the same types of plots for a star-shaped graph with fixed maximum node degree $d = 40$, and Figure 6 shows the analogous plots for a grid graph with fixed degree $d = 4$. As in the chain case, these plots show the same type of stacking effect in terms of the scaled sample size $n/\log p$, when the degree d and other parameters $((\alpha, K_{\Gamma^*}, K_{\Sigma^*}))$ are held fixed.

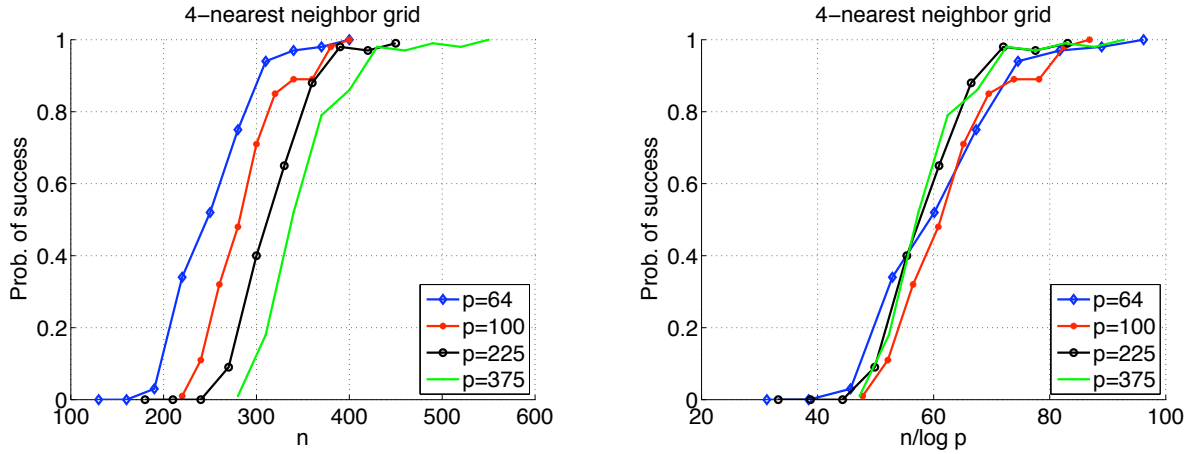
5.2 Dependence on the maximum node degree

Panel (a) of Figure 7 plots the probability of correct signed edge-set recovery against the sample size n for star-shaped graphs; each curve corresponds to a different choice of maximum node degree d, allowing us to investigate the dependence of the sample size on this parameter. So as to control these comparisons, the models are chosen such that quantities other than the maximum node-degree d are fixed: in particular, we fix the number of nodes $p = 200$, and the edge covariance entries are set as $\Sigma^*_{ij} = 2.5/d$ for $(i, j) \in E$ so that the quantities $(K_{\Sigma^*}, K_{\Gamma^*}, \alpha)$ remain constant. The minimum value $\theta_{\min}$ in turn scales as $1/d$. Observe how the plots in panel (a) shift to the right as the maximum node degree d is increased, showing that star-shaped graphs with higher degrees are more difficult. In panel (b) of Figure 7, we plot the same data versus the rescaled sample size n/d. Recall that if all the curves were to stack up under this rescaling, then it means the required sample size n scales linearly with d. These plots are closer to aligning than the unrescaled plots, but the agreement is not perfect. In particular, observe that the curve d (right-most in panel (a)) remains a bit to the right in panel (b), which suggests that a somewhat more aggressive rescaling—perhaps n/d^γ for some $\gamma \in (1, 2)$—is appropriate.

Note that for $\theta_{\min}$ scaling as $1/d$, the sufficient condition from Theorem 2, as summarized in equation (71), is $n = \Omega(d^2 \log p)$, which appears to be overly conservative based on these data. Thus, it might be possible to tighten our theory under certain regimes.

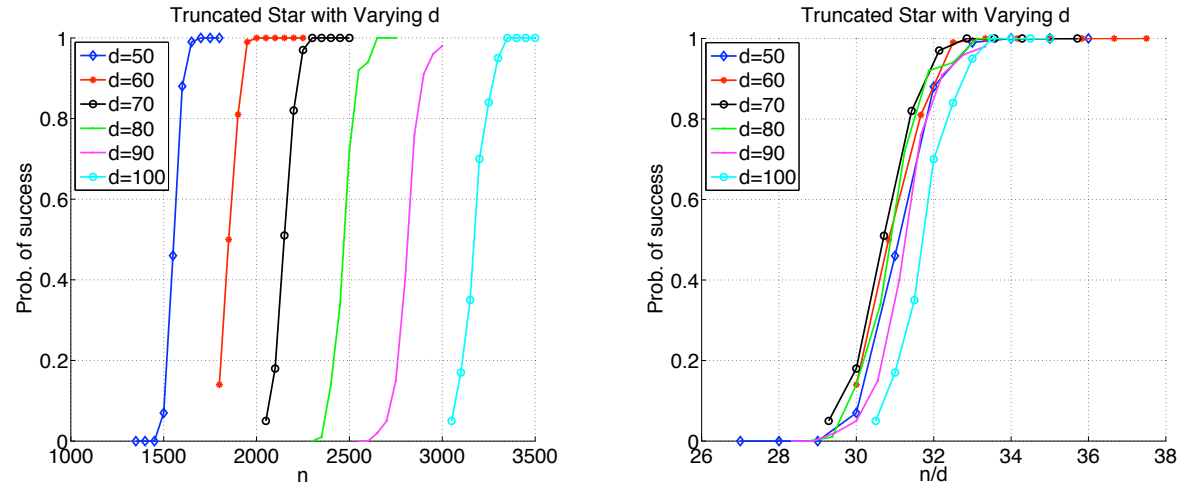
5.3 Dependence on covariance and Hessian terms

Next, we study the dependence of the sample size required for model selection consistency on the model complexity term K defined in (70), which is a collection of the quantities K_{Σ^*}, K_{Γ^*} and α defined by the covariance matrix and Hessian, as well as the minimum value $\theta_{\min}$. Figure 8 plots the probability of correct signed edge-set recovery versus the sample size n for chain graphs. Here each curve corresponds to a different setting of the model complexity factor



(a) (b)

Figure 6. Simulations for 2-dimensional lattice with 4-nearest-neighbor interaction, edge strength interactions $\Theta^*_{ij} = 0.1$, and a varying number of nodes p. Plots of probability of correct signed edge-set recovery versus the sample size n in panel (a), and versus the rescaled sample size $n/\log p$ in panel (b). Each point corresponds to the average over $N = 100$ trials.



(a) (b)

Figure 7. Simulations for star graphs with fixed number of nodes $p = 200$, varying maximal (hub) degree d, edge covariances $\Sigma^*_{ij} = 2.5/d$. Plots of probability of correct signed edge-set recovery versus the sample size n in panel (a), and versus the rescaled sample size n/d in panel (b).

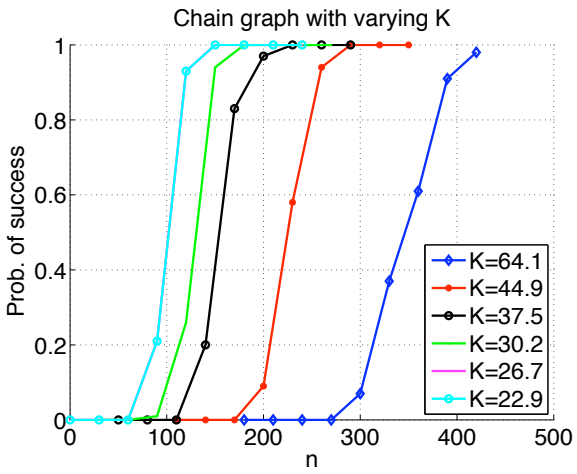


Figure 8. Simulations for chain graph with fixed number of nodes $p = 120$, and varying model complexity K. Plot of probability of correct signed edge-set recovery versus the sample size n.

K, but with a fixed number of nodes $p = 120$, and maximum node-degree $d = 2$. We varied the actor K by varying the value ρ of the edge covariances $\Sigma_{ij} = \rho$, $(i, j) \in E$. Notice how the curves, each of which corresponds to a different model complexity factor, shift rightwards as K is increased so that models with larger values of K require greater number of samples n to achieve the same probability of correct model selection. These rightward-shifts are in qualitative agreement with the prediction of Theorem 1, but we suspect that our analysis is not sharp enough to make accurate quantitative predictions regarding this scaling.

5.4 Convergence rates in elementwise ℓ_∞-norm

Finally, we report some simulation results on the convergence rate in elementwise ℓ_∞-norm. According to Corollary 1, in the case of sub-Gaussian tails, if the elementwise ℓ_∞-norm should decay at rate $\mathcal{O}(\sqrt{\frac{\log p}{n}})$, once the sample size n is sufficiently large. Figure 9 shows the behavior of the elementwise ℓ_∞-norm for star-shaped graphs of varying sizes p. The results reported here correspond to the maximum degree $d = \lceil 0.1p \rceil$; we also performed analogous experiments for $d = \mathcal{O}(\log p)$ and $d = \mathcal{O}(1)$, and observed qualitatively similar behavior. The edge correlations were set as $\Sigma^*_{ij} = 2.5/d$ for all $(i, j) \in E$ so that the quantities $(K_{\Sigma^*}, K_{\Gamma^*}, \alpha)$ remain constant. With these settings, each curve in Figure 9 corresponds to a different problem size, and plots the elementwise ℓ_∞-error versus the rescaled sample size $n/\log p$, so that we expect to see curves of the form $f(t) = 1/\sqrt{t}$. The curves show that when the rescaled sample size $(n/\log p)$ is larger than some threshold (roughly 40 in the plots shown), the elementwise ℓ_∞ norm decays at the rate $\sqrt{\frac{\log p}{n}}$, which is consistent with Corollary 1.

6 Discussion

The focus of this paper is the analysis of the high-dimensional scaling of the ℓ_1-regularized log determinant problem (11) as an estimator of the concentration matrix of a random vector. Our main contributions were to derive sufficient conditions for its model selection consistency as well as convergence rates in both elementwise ℓ_∞-norm, as well as Frobenius and spectral norms. Our results allow for a range of tail behavior, ranging from the exponential-type decay provided by Gaussian random vectors (and sub-Gaussian more generally), to polynomial-type decay guaranteed by moment conditions. In the Gaussian case, our results have natural interpretations in terms of Gaussian Markov random fields.

Our main results relate the i.i.d. sample size n to various parameters of the problem required to achieve consistency. In addition to the dependence on matrix size p, number of edges s and graph degree d, our analysis also illustrates the

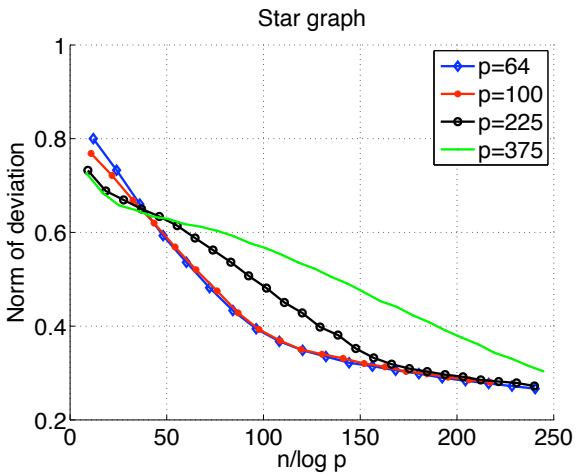


Figure 9. Simulations for a star graph with varying number of nodes p, maximum node degree $d = \lceil 0.1p \rceil$, edge covariances $\Sigma^*_{ij} = 2.5/d$. Plot of the element-wise ℓ_∞ norm of the concentration matrix estimate error $\|\widehat{\Theta} - \Theta^*\|_\infty$ versus the rescaled sample size $n/\log(p)$.

role of other quantities, related to the structure of the covariance matrix Σ^* and the Hessian of the objective function, that have an influence on consistency rates. Our main assumption is an irrepresentability or mutual incoherence condition, similar to that required for model selection consistency of the Lasso, but involving the Hessian of the log-determinant objective function (11), evaluated at the true model. When the distribution of X is multivariate Gaussian, this Hessian is the Fisher information matrix of the model, and thus can be viewed as an edge-based counterpart to the usual node-based covariance matrix We report some examples where irrepresentability condition for the Lasso hold and the log-determinant condition fails, but we do not know in general if one requirement dominates the other. In addition to these theoretical results, we provided a number of simulation studies showing how the sample size required for consistency scales with problem size, node degrees, and the other complexity parameters identified in our analysis.

There are various interesting questions and possible extensions to this paper. First, in the current paper, we have only derived sufficient conditions for model selection consistency. As in past work on the Lasso [23], it would also be interesting to derive a *converse result*—namely, to prove that if the sample size n is smaller than some function of (p, d, s) and other complexity parameters, then regardless of the choice of regularization constant, the log-determinant method fails to recover the correct graph structure. Second, while this paper studies the problem of estimating a fixed graph or concentration matrix, a natural extension would allow the graph to vary over time, a problem setting which includes the case where the observations are dependent. For instance, Zhou et al. [27] study the estimation of the covariance matrix of a Gaussian distribution in a time-varying setting, and it would be interesting to extend results of this paper to this more general setting.

Acknowledgements

We thank Shuheng Zhou for helpful comments on an earlier draft of this work. Work was partially supported by NSF grant DMS-0605165. Yu also acknowledges support from ARO W911NF-05-1-0104, NSFC-60628102, and a grant from MSRA.

A Proof of Lemma 3

In this appendix, we show that the regularized log-determinant program (11) has a unique solution whenever $\lambda_n > 0$, and the diagonal of the sample covariance $\widehat{\Sigma}^n$ is strictly positive. By the strict convexity of the log-determinant barrier [3], if the minimum is attained, then it is unique, so that it remains to show that the minimum is achieved. If

$\lambda_n > 0$, then by Lagrangian duality, the problem can be written in an equivalent constrained form:

$$\min_{\Theta \succ 0, \|\Theta\|_{1,\text{off}} \leq C(\lambda_n)} \left\{ \langle\langle \Theta, \ \widehat{\Sigma}^n \rangle\rangle - \log\det(\Theta) \right\} \tag{72}$$

for some $C(\lambda_n) < +\infty$. Since the off-diagonal elements remain bounded within the ℓ_1-ball, the only possible issue is the behavior of the objective function for sequences with possibly unbounded diagonal entries. Since any Θ in the constraint set is positive-definite, its diagonal entries are positive, and hence bounded below by zero. Further, by Hadamard's inequality for positive definite matrices [12], we have $\log\det\Theta \leq \sum_{i=1}^p \log\Theta_{ii}$, so that

$$\sum_{i=1}^p \Theta_{ii}\widehat{\Sigma}^n_{ii} - \log\det\Theta \quad \geq \quad \sum_{i=1}^p \left\{ \Theta_{ii}\widehat{\Sigma}^n_{ii} - \log\Theta_{ii} \right\}.$$

As long as $\widehat{\Sigma}^n_{ii} > 0$ for each $i = 1, \ldots, p$, this function is coercive, meaning that it diverges to infinity for any sequence $\|(\Theta^t_{11}, \ldots, \Theta^t_{pp})\| \to +\infty$. Consequently, the minimum is attained.

Returning to the penalized form (11), by standard optimality conditions for convex programs, a matrix $\widehat{\Sigma} \succ 0$ is optimal if and only 0 belongs to the sub-differential of the objective, or equivalently if and only if there exists a matrix $\widehat{Z}$ in the sub-differential of the off-diagonal norm $\|\cdot\|_{1,\text{off}}$ such that

$$\widehat{\Sigma} - \widehat{\Theta}^{-1} + \lambda\widehat{Z} = 0,$$

as claimed.

B Proof of Lemma 5

We write the remainder in the form

$$R(\Delta) \quad = \quad (\Theta^* + \Delta)^{-1} - \Theta^{*-1} + \Theta^{*-1}\Delta\Theta^{*-1}.$$

By sub-multiplicativity of the $|\!|\!|\cdot|\!|\!|_\infty$ matrix norm, for any two $p \times p$ matrices A, B, we have $|\!|\!|AB|\!|\!|_\infty \leq |\!|\!|A|\!|\!|_\infty |\!|\!|B|\!|\!|_\infty$, so that

$$\begin{aligned} |\!|\!|\Theta^{*-1}\Delta|\!|\!|_\infty \quad &\leq \quad |\!|\!|\Theta^{*-1}|\!|\!|_\infty |\!|\!|\Delta|\!|\!|_\infty \\ &\leq \quad K_{\Sigma^*}\, d\|\Delta\|_\infty \ < \ 1/3, \end{aligned} \tag{73}$$

where we have used the definition of K_{Σ^*}, the fact that Δ has at most d non-zeros per row/column, and our assumption $\|\Delta\|_\infty \ < 1/(3K_{\Sigma^*})$. Consequently, we have the convergent matrix expansion

$$\begin{aligned} \big(\Theta^* + \Delta\big)^{-1} \quad &= \quad \big(\Theta^*\big(I + \Theta^{*-1}\Delta\big)\big)^{-1} \\ &= \quad \big(I + \Theta^{*-1}\Delta\big)^{-1}\big(\Theta^*\big)^{-1} \\ &= \quad \sum_{k=0}^{\infty}(-1)^k\big(\Theta^{*-1}\Delta\big)^k\big(\Theta^*\big)^{-1} \\ &= \quad \Theta^{*-1} - \Theta^{*-1}\Delta\Theta^{*-1} + \sum_{k=2}^{\infty}(-1)^k\big(\Theta^{*-1}\Delta\big)^k\big(\Theta^*\big)^{-1} \\ &= \quad \Theta^{*-1} - \Theta^{*-1}\Delta\Theta^{*-1} + \Theta^{*-1}\Delta\Theta^{*-1}\Delta J\Theta^{*-1}, \end{aligned}$$

where $J = \sum_{k=0}^{\infty}(-1)^k\big(\Theta^{*-1}\Delta\big)^k$.

We now prove the bound (58) on the remainder as follows. Let e_i denote the unit vector with 1 in position i and zeroes elsewhere. From equation (57), we have

$$\begin{aligned} \|R(\Delta)\|_\infty \quad &= \quad \max_{i,j} |e_i^T\Theta^{*-1}\Delta\ \Theta^{*-1}\Delta J\Theta^{*-1}e_j| \\ &\leq \quad \max_i \|e_i^T\Theta^{*-1}\Delta\|_\infty \ \max_j \|\Theta^{*-1}\Delta J\Theta^{*-1}e_j\|_1, \end{aligned}$$

which follows from the fact that for any vectors $a, b \in \mathbb{R}^p$, $|a^T b| \leq \|a\|_\infty \|b\|_1$. This in turn can be simplified as,

$$\|R(\Delta)\|_\infty \quad \leq \quad \max_i \|e_i^T \Theta^{*-1}\|_1 \; \|\Delta\|_\infty \; \max_j \|\Theta^{*-1} \Delta J \Theta^{*-1} e_j\|_1$$

since for any vector $u \in \mathbb{R}^p$, $\|u^T \Delta\|_\infty \leq \|u\|_1 \|\Delta\|_\infty$, where $\|\Delta\|_\infty$ is the elementwise ℓ_∞-norm. Continuing on, we have

$$\|R(\Delta)\|_\infty \quad \leq \quad |\!|\!|\Theta^{*-1}|\!|\!|_\infty \; \|\Delta\|_\infty |\!|\!| \; |\!|\!|\Theta^{*-1} \Delta J \Theta^{*-1}|\!|\!|_1,$$

where $|\!|\!|A|\!|\!|_1 := \max_{\|x\|_1 = 1} \|Ax\|_1$ is the ℓ_1-operator norm. Since $|\!|\!|A|\!|\!|_1 = |\!|\!|A^T|\!|\!|_\infty$, we have

$$\begin{aligned} \|R(\Delta)\|_\infty &\leq \|\Delta\|_\infty |\!|\!|\Theta^{*-1}|\!|\!|_\infty \; |\!|\!|\Theta^{*-1} J^T \Delta \Theta^{*-1}|\!|\!|_\infty \\ &\leq \|\Delta\|_\infty \; K_{\Sigma^*} |\!|\!|\Theta^{*-1}|\!|\!|_\infty^2 |\!|\!| J^T |\!|\!|_\infty |\!|\!| \Delta |\!|\!|_\infty \end{aligned} \tag{74}$$

Recall that $J = \sum_{k=0}^\infty (-1)^k \big(\Theta^{*-1} \Delta\big)^k$. By sub-multiplicativity of $|\!|\!| \cdot |\!|\!|_\infty$ matrix norm, we have

$$|\!|\!| J^T |\!|\!|_\infty \leq \sum_{k=0}^\infty |\!|\!| \Delta \Theta^{*-1} |\!|\!|_\infty^k \leq \frac{1}{1 - |\!|\!|\Theta^{*-1}|\!|\!|_\infty |\!|\!| \Delta |\!|\!|_\infty} \leq \frac{3}{2},$$

since $|\!|\!|\Theta^{*-1}|\!|\!|_\infty |\!|\!| \Delta |\!|\!|_\infty < 1/3$ from equation (73). Substituting this in (74), we obtain

$$\begin{aligned} \|R(\Delta)\|_\infty &\leq \frac{3}{2} \|\Delta\|_\infty \; K_{\Sigma^*} |\!|\!|\Theta^{*-1}|\!|\!|_\infty^2 |\!|\!| \Delta |\!|\!|_\infty \\ &\leq \frac{3}{2} d \|\Delta\|_\infty^2 \; K_{\Sigma^*}^3, \end{aligned}$$

where the final line follows since $|\!|\!| \Delta |\!|\!|_\infty \leq d \|\Delta\|_\infty$, and since Δ has at most d non-zeroes per row/column.

C Proof of Lemma 6

By following the same argument as in Appendix A, we conclude that the restricted problem (49) has a unique optimum $\widetilde{\Theta}$. Let $\widetilde{Z}$ be any member of the sub-differential of $\|\cdot\|_{1,\text{off}}$, evaluated at $\widetilde{\Theta}$. By Lagrangian theory, the witness $\widetilde{\Theta}$ must be an optimum of the associated Lagrangian problem

$$\min_{\Theta \succ 0, \; \Theta_{S^c} = 0} \big\{ \langle\!\langle \Theta, \; \widehat{\Sigma} \rangle\!\rangle - \log\det(\Theta) + \lambda_n \langle\!\langle \Theta, \; \widetilde{Z} \rangle\!\rangle \big\}.$$

In fact, since this Lagrangian is strictly convex, $\widetilde{\Theta}$ is the only optimum of this problem. Since the log-determinant barrier diverges as Θ approaches the boundary of the positive semi-definite cone, we must have $\widetilde{\Theta} \succ 0$. If we take partial derivatives of the Lagrangian with respect to the unconstrained elements Θ_S, these partial derivatives must vanish at the optimum, meaning that we have the zero-gradient condition

$$G(\Theta_S) \quad = \quad -\Theta_S^{-1} + \widehat{\Sigma}_S + \lambda_n \widetilde{Z}_S \; = \; 0. \tag{75}$$

To be clear, Θ is the $p \times p$ matrix with entries in S equal to Θ_S and entries in S^c equal to zero. Since this zero-gradient condition is necessary and sufficient for an optimum of the Lagrangian problem, it has a unique solution (namely, $\widetilde{\Theta}_S$).

Our goal is to bound the deviation of this solution from Θ_S^*, or equivalently to bound the deviation $\Delta = \widetilde{\Theta} - \Theta^*$. Our strategy is to show the existence of a solution Δ to the zero-gradient condition (75) that is contained inside the ball $\mathbb{B}(r)$ defined in equation (61). By uniqueness of the optimal solution, we can thus conclude that $\widetilde{\Theta} - \Theta^*$ belongs this ball. In terms of the vector $\overline{\Delta}_S = \overline{\Theta}_S - \overline{\Theta^*}_S$, let us define a map $F : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ via

$$F(\overline{\Delta}_S) \quad := \quad -\big(\Gamma_{SS}^*\big)^{-1} \big(\overline{G}(\Theta_S^* + \Delta_S)\big) + \overline{\Delta}_S, \tag{76}$$

where $\overline{G}$ denotes the vectorized form of G. Note that by construction, $F(\overline{\Delta}_S) = \overline{\Delta}_S$ holds if and only if $G(\Theta^*_S + \Delta_S) = G(\Theta_S) = 0$.

We now claim that $F(\mathbb{B}(r)) \subseteq \mathbb{B}(r)$. Since F is continuous and $\mathbb{B}(r)$ is convex and compact, this inclusion implies, by Brouwer's fixed point theorem [20], that there exists some fixed point $\overline{\Delta}_S \in \mathbb{B}(r)$. By uniqueness of the zero gradient condition (and hence fixed points of F), we can thereby conclude that $\|\widetilde{\Theta}_S - \Theta^*_S\|_\infty \leq r$.

Let $\Delta \in \mathbb{R}^{p \times p}$ denote the zero-padded matrix, equal to Δ_S on S and zero on S^c. By definition, we have

$$\begin{aligned} G(\Theta^*_S + \Delta_S) &= -(\Theta^* + \Delta)^{-1}_S + \widehat{\Sigma}_S + \lambda_n \widetilde{Z}_S \\ &= \big[-(\Theta^* + \Delta)^{-1}_S + \Theta^*_S \big] + \big[\widehat{\Sigma}_S - (\Theta^*)^{-1}_S \big] + \lambda_n \widetilde{Z}_S \\ &= \big[-(\Theta^* + \Delta)^{-1}_S + \Theta^*_S \big] + W_S + \lambda_n \widetilde{Z}_S, \end{aligned} \tag{77}$$

where we have used the definition $W = \widehat{\Sigma} - \Sigma^*$.

For any $\Delta_S \in \mathbb{B}(r)$, we have

$$\begin{aligned} |\!|\!| \Theta^{*-1} \Delta |\!|\!|_\infty &\leq |\!|\!| \Theta^{*-1} |\!|\!|_\infty |\!|\!| \Delta |\!|\!|_\infty \\ &\leq K_{\Sigma^*} \, d \|\Delta\|_\infty, \end{aligned} \tag{78}$$

where $\|\Delta\|_\infty$ denotes the elementwise ℓ_∞-norm (as opposed to the ℓ_∞-operator norm $|\!|\!| \Delta |\!|\!|_\infty$), and the inequality follows since Δ has at most d non-zero entries per row/column,

By the definition (61) of the radius r, and the assumed upper bound (59), we have $\|\Delta\|_\infty \leq r \leq \frac{1}{3K_{\Sigma^*} d}$, so that the results of Lemma 5 apply. By using the definition (52) of the remainder, taking the vectorized form of the expansion (57), and restricting to entries in S, we obtain the expansion

$$\operatorname{vec}\left(\left(\Theta^* + \Delta\right)^{-1} - \Theta^{*-1} \right)_S + \Gamma^*_{SS} \overline{\Delta}_S \;=\; \operatorname{vec}\left((\Theta^{*-1} \Delta)^2 J \Theta^{*-1} \right)_S. \tag{79}$$

Using this expansion (79) combined with the expression (77) for G, we have

$$\begin{aligned} F(\overline{\Delta}_S) &= -\left(\Gamma^*_{SS}\right)^{-1} \overline{G}(\Theta^*_S + \Delta_S) + \overline{\Delta}_S \\ &= \left(\Gamma^*_{SS}\right)^{-1} \operatorname{vec}\left\{ \left[\left(\Theta^* + \Delta\right)^{-1} - \Theta^{*-1} \right]_S - W_S - \lambda_n \widetilde{Z}_S \right\} + \overline{\Delta}_S \\ &= \underbrace{\left(\Gamma^*_{SS}\right)^{-1} \operatorname{vec}\left[(\Theta^{*-1} \Delta)^2 J \Theta^{*-1} \right]_S}_{T_1} - \underbrace{\left(\Gamma^*_{SS}\right)^{-1} \left(\overline{W}_S + \lambda_n \overline{\widetilde{Z}}_S \right)}_{T_2}. \end{aligned}$$

The second term is easy to deal with: using the definition $K_{\Gamma^*} = |\!|\!| (\Gamma^*_{SS})^{-1} |\!|\!|_\infty$, we have $\|T_2\|_\infty \leq K_{\Gamma^*} \left(\|W\|_\infty + \lambda_n \right) = r/2$. It now remains to show that $\|T_1\|_\infty \leq r/2$. We have

$$\begin{aligned} \|T_1\|_\infty &\leq K_{\Gamma^*} \left\| \operatorname{vec}\left[(\Theta^{*-1} \Delta)^2 J \Theta^{*-1} \right]_S \right\|_\infty \\ &\leq K_{\Gamma^*} \|R(\Delta)\|_\infty, \end{aligned}$$

where we used the expanded form (57) of the remainder, Applying the bound (58) from Lemma 5, we obtain

$$\|T_1\|_\infty \;\leq\; \frac{3}{2} d K^3_{\Sigma^*} K_{\Gamma^*} \, \|\Delta\|^2_\infty \;\leq\; \frac{3}{2} d K^3_{\Sigma^*} K_{\Gamma^*} \, r^2.$$

Since $r \leq \frac{1}{3K^3_{\Sigma^*} K_{\Gamma^*} d}$ by assumption (59), we conclude that

$$\|T_1\|_\infty \;\leq\; \frac{3}{2} d K^3_{\Sigma^*} K_{\Gamma^*} \, \frac{1}{3K^3_{\Sigma^*} K_{\Gamma^*} d} r \;=\; r/2,$$

thereby establishing the claim.

D Proof of Lemma 1

For each pair (i,j) and $\nu > 0$, define the event

$$\mathbb{A}_{ij}(\nu) \quad := \quad \{|\frac{1}{n}\sum_{k=1}^{n} X_i^{(k)} X_j^{(k)} - \Sigma^*_{ij}| > \nu\}.$$

As the sub-Gaussian assumption is imposed on the variables $\{X_i^{(k)}\}$ directly, as in Lemma A.3 of Bickel and Levina [2], our proof proceeds by first decoupling the products $X_i^{(k)} X_j^{(k)}$. For each pair (i,j), we define $\rho^*_{ij} = \Sigma^*_{ij}/\sqrt{\Sigma^*_{ii}\Sigma^*_{jj}}$, and the rescaled random variables $\overline{X}_i^{(k)} := X_{ij}^{(k)}/\sqrt{\Sigma^*_{ii}}$. Noting that the strict positive definiteness of Σ^* implies that $|\rho^*_{ij}| < 1$, we can also define the auxiliary random variables

$$U_{ij}^{(k)} := \overline{X}_i^{(k)} + \overline{X}_j^{(k)} \quad \text{and} \quad V_{ij}^{(k)} := \overline{X}_i^{(k)} - \overline{X}_j^{(k)}. \tag{80}$$

With this notation, we then claim:

Lemma 9. *Suppose that each $\overline{X}_i^{(k)}$ is sub-Gaussian with parameter σ. Then for each node pair (i,j), the following properties hold:*

(a) For all $k = 1, \ldots, n$, the random variables $U_{ij}^{(k)}$ and $V_{ij}^{(k)}$ are sub-Gaussian with parameters 2σ.

(b) For all $\nu > 0$, the probability $\mathbb{P}[\mathbb{A}_{ij}(\nu)]$ is upper bounded by

$$\mathbb{P}\big[|\sum_{k=1}^{n}(U_{ij}^{(k)})^2 - 2(1-\rho^*_{ij})| > \frac{2n\nu}{\sqrt{\Sigma^*_{ii}\Sigma^*_{jj}}}\big] + \mathbb{P}\big[|\sum_{k=1}^{n}(V_{ij}^{(k)})^2 - 2(1-\rho^*_{ij})| > \frac{2n\nu}{\sqrt{\Sigma^*_{ii}\Sigma^*_{jj}}}\big].$$

Proof. (a) For any $r \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(r\, U_{ij}^{(k)})] \quad = \quad \mathbb{E}\Big[\exp\big(r\,\overline{X}_i^{(k)}\big)\ \exp\big(r\overline{X}_j^{(k)}\big)\Big] \ \leq\ \mathbb{E}\Big[\exp\big(2r\,\overline{X}_i^{(k)}\big)\Big]^{1/2}\ \Big[\exp\big(2r\overline{X}_j^{(k)}\big)\Big]^{1/2},$$

where we have used the Cauchy-Schwarz inequality. Since the variables $\overline{X}_i^{(k)}$ and $\overline{X}_j^{(k)}$ are sub-Gaussian with parameter σ, we have

$$\mathbb{E}\Big[\exp\big(2r\,\overline{X}_i^{(k)}\big)\Big]^{1/2}\ \mathbb{E}\Big[\exp\big(2r\overline{X}_j^{(k)}\big)\Big]^{1/2} \quad \leq \quad \exp(\sigma^2\frac{r^2}{2})\ \exp(\sigma^2\frac{r^2}{2}),$$

so that $U_{ij}^{(k)}$ is sub-Gaussian with parameter 2σ as claimed.
(b) By straightforward algebra, we have the decomposition

$$\sum_{k=1}^{n}(\overline{X}_i^{(k)}\overline{X}_j^{(k)} - \rho^*_{ij}) \ = \ \{\frac{1}{4}\sum_{i=1}^{n}\{(\overline{X}_i^{(k)} + \overline{X}_j^{(k)})^2 - 2(1+\rho^*_{st})\} \ - \ \{\frac{1}{4}\sum_{i=1}^{n}\{(X_s^{*(k)} - X_t^{*(k)})^2 - 2(1-\rho^*_{st})\}.$$

By union bound, we obtain that $\mathbb{P}[\mathbb{A}_{ij}(\nu)]$ is upper bounded by

$$\mathbb{P}\Big[|\sum_{k=1}^{n}(U_{ij}^{(k)})^2 - 2\,(1+\rho^*_{ij})| \geq \frac{4n\nu}{2\,\sqrt{\Sigma^*_{ii}\Sigma^*_{jj}}}\Big] + \mathbb{P}\Big[|\sum_{k=1}^{n}(V_{ij}^{(k)})^2 - 2\,(1-\rho^*_{ij})| \geq \frac{4n\nu}{2\,\sqrt{\Sigma^*_{ii}\Sigma^*_{jj}}}\Big], \tag{81}$$

which completes the proof of Lemma 9(b). □

It remains to control the terms $\sum_{k=1}^{n}(U_{ij}^{(k)})^2$ and $\sum_{k=1}^{n}(V_{ij}^{(k)})^2$. We do so by exploiting tail bounds [6] for sub-exponential random variables. A zero-mean random variable Z is said to be *sub-exponential* if there exists a constant $\gamma \in (0,\infty)$ and $\phi \in (0,\infty]$ such that

$$\mathbb{E}[\exp(tZ)] \quad \leq \quad \exp(\gamma^2\, t^2/2) \qquad \text{for all } t \in (-\phi, \phi). \tag{82}$$

Note that for $\phi < +\infty$, this requirement is a weakening of sub-Gaussianity, since the inequality is only required to hold on the interval $(-\phi, +\phi)$.

Now consider the variates $Z_{k;ij} := (U_{ij}^{(k)})^2 - 2\,(1+\rho_{ij}^*)$. Note that they are zero-mean; we also claim they are sub-exponential.

Lemma 10. *For all $k \in \{1,\ldots,n\}$ and node-pairs $(i,j) \in V \times V$, the variables*

$$Z_{k;ij} \quad := \quad (U_i^{(k)})^2 - 2\,(1+\rho_{ij}^*)$$

are sub-exponential with parameter $\gamma_U = 16(1+4\sigma^2)$ in the interval $(-\phi_U, \phi_U)$, with $\phi_U = 1/(16(1+4\sigma^2))$.

We can exploit this lemma to apply tail bounds for sums of i.i.d. sub-exponential variates (Thm. 5.1, [6]). Doing so yields that for $t \leq \gamma_U^2 \phi_U$, we have $\mathbb{P}\Big[|\sum_{k=1}^n (U_{ij}^{(k)})^2 - 2(1+\rho_{ij}^*)| > nt\Big] \leq 2\exp\big\{-nt^2/(2\gamma_U^2)\big\}$. Setting $t = 2\nu/\max_i \Sigma_{ii}^*$, and noting that $(2n\nu/\sqrt{\Sigma_{ii}^*\Sigma_{jj}^*}) \geq (2n\nu/\max_i \Sigma_{ii}^*)$, we obtain

$$\begin{aligned}
\mathbb{P}\Big[|\sum_{k=1}^n (U_{ij}^{(k)})^2 - 2(1+\rho_{ij}^*)| > \frac{2n\nu}{\sqrt{\Sigma_{ii}^*\Sigma_{jj}^*}}\Big] \quad &\leq \quad 2\exp\Big\{-\frac{2n\nu^2}{\max_i(\Sigma_{ii}^*)^2\,\gamma_U^2}\Big\} \\
&\leq \quad 2\exp\Big\{-\frac{n\nu^2}{\max_i(\Sigma_{ii}^*)^2\,128\,(1+4\sigma^2)^2}\Big\},
\end{aligned}$$

for all $\nu \leq 8(\max_i \Sigma_{ii}^*)\,(1+4\sigma^2)$. A similar argument yields the same tail bound for the deviation involving $V_{ij}^{(k)}$. Consequently, using Lemma 9(b), we conclude that

$$\mathbb{P}[\mathbb{A}_{ij}(\nu)] \quad \leq \quad 4\exp\big\{-\frac{n\nu^2}{\max_i(\Sigma_{ii}^*)^2\,128\,(1+4\sigma^2)^2}\big\}.,$$

valid for $\nu \leq 8(\max_i \Sigma_{ii}^*)\,(1+4\sigma^2)$, as required. It only remains to prove Lemma 10.

Proof of Lemma 10. If we can obtain a bound $B > 0$ such that

$$\sup_{m\geq 2}\Big[\frac{\mathbb{E}(|Z_{k;ij}|^m}{m!}\Big]^{1/m} \quad \leq \quad B,$$

it then follows (Thm. 3.2, [6]) that $Z_{k;ij}$ is sub-exponential with parameter $2B$ in the interval $(-\frac{1}{2B}, +\frac{1}{2B})$. We obtain such a bound B as follows. Using the inequality $(a+b)^m \leq 2^m(a^m+b^m)$, valid for any real numbers a, b, we have

$$\mathbb{E}(|Z_{k;ij}|^m) \leq 2^m\,\Big(\mathbb{E}(|U_{ij}^{(k)}|^{2m}) + (2(1+\rho_{ij}^*))^m\Big). \tag{83}$$

Recalling that $U_{ij}^{(k)}$ is sub-Gaussian with parameter 2σ, from Lemma 1.4 of Buldygin and Kozachenko [6] regarding the moments of sub-Gaussian variates, we have $\mathbb{E}[|U_{ij}^{(k)}|^{2m}] \leq 2(2m/e)^m(2\sigma)^{2m}$. Thus, noting the inequality $m! \geq (m/e)^m$, it follows that $\mathbb{E}[|U_{ij}^{(k)}|^{2m}]/m! \leq 2^{3m+1}\sigma^{2m}$. It then follows from equation (83) that

$$\begin{aligned}
\Big[\frac{\mathbb{E}(|Z_{k;ij}|^m)}{m!}\Big]^{1/m} \quad &\leq \quad 2^{1/m}\left((2^{4m+1}\sigma^{2m})^{1/m} + \frac{4(1+\rho_{ij}^*)}{(m!)^{1/m}}\right) \\
&\leq \quad 2^{1/m}\left(2^{1/m}\,16\,\sigma^2 + \frac{4(1+\rho_{ij}^*)}{(m!)^{1/m}}\right),
\end{aligned}$$

where we have used the inequality $(x+y)^{1/m} \leq 2^{1/m}(x^{1/m}+y^{1/m})$, valid for any integer $m \in \mathbb{N}$ and real numbers $x, y > 0$. Since the bound is a decreasing function of m, it follows that

$$\begin{aligned}
\sup_{m\geq 2}\Big[\frac{\mathbb{E}(|Z_{k;ij}|^m}{m!}\Big]^{1/m} \quad &\leq \quad 2^{1/2}\left(2^{1/2}\,16\,\sigma^2 + \frac{4(1+\rho_{ij}^*)}{(2)^{1/2}}\right) \\
&\leq \quad 32\sigma^2 + 8 = 8(1+4\sigma^2),
\end{aligned}$$

where we have used the fact that $|\rho_{ij}^*| \leq 1$. The claim of the lemma thus follows. □

E Proof of Lemma 2

Define the random variables $W_{ij}^{(k)} = X_i^{(k)} X_j^{(k)} - \Sigma_{ij}^*$, and note that they have mean zero. By applying the Chebyshev inequality, we obtain

$$\begin{aligned} \mathbb{P}\big[\Big|\sum_{k=1}^n W_{ij}^{(k)}\Big| > n\,\nu\big] &= \mathbb{P}\big[(|\sum_{k=1}^n W_{ij}^{(k)}\big)^{2m} > (n\,\nu)^{2m}\big] \\ &\leq \frac{\mathbb{E}\big[\big(\sum_{k=1}^n W_{ij}^{(k)}\big)^{2m}\big]}{n^{2m}\,\nu^{2m}}. \end{aligned} \tag{84}$$

Letting $\mathcal{A} = \{(a_1, \ldots, a_n) \mid a_i \in \{0, \ldots, 2m\}, \sum_{i=1}^n a_i = 2m\}$, by the multinomial theorem, we have

$$\begin{aligned} \mathbb{E}\big[\big(\sum_{k=1}^n W_{ij}^{(k)}\big)^{2m}\big] &= \mathbb{E}\Big[\sum_{\mathbf{a}\in\mathcal{A}} \binom{2m}{a_1, \ldots, a_n} \prod_{k=1}^n (W_{ij}^{(k)})^{a_k}\Big] \\ &= \sum_{a\in\mathcal{A}} \binom{2m}{a_1, \ldots, a_n} \prod_{k=1}^n \mathbb{E}\big[(W_{ij}^{(k)})^{a_k}\big], \end{aligned}$$

where the final equality uses linearity of expectation, and the independence of the variables $\{W_{ij}^{(k)}\}_{k=1}^n$.

Since the variables $W_{ij}^{(k)}$ are zero-mean, the product $\prod_{k=1}^n \mathbb{E}\big[(W_{ij}^{(k)})^{a_k}\big]$ vanishes for any multi-index $a \in \mathcal{A}$ such that $a_k = 1$ for at least one k. Accordingly, defining the subset

$$\mathcal{A}_{-1} := \{(a_1, \ldots, a_n) \mid a_i \in \{0, 2, \ldots, 2m\}, \ \sum_{i=1}^n a_i = 2m\},$$

we have

$$\begin{aligned} \mathbb{E}\big[\big(\sum_{k=1}^n W_{ij}^{(k)}\big)^{2m}\big] &= \sum_{a\in\mathcal{A}_{-1}} \binom{2m}{a_1, \ldots, a_n} \prod_{k=1}^n \mathbb{E}\big[(W_{ij}^{(k)})^{a_k}\big] \\ &\leq \underbrace{\Big(\sum_{a\in\mathcal{A}_{-1}} \binom{2m}{a_1, \ldots, a_n}\Big)}_{T_1} \ \underbrace{\max_{a\in\mathcal{A}_{-1}} \prod_{k=1}^n \mathbb{E}\big[(W_{ij}^{(k)})^{a_k}\big]}_{T_2}. \end{aligned} \tag{85}$$

The quantity T_1 is equal to the number of ways to put $2m$ balls in n bins such that if a bin contains a ball, it should have at least two balls. Note that this implies there can then be at most m bins containing a ball. Consequently, the term T_1 is bounded above by the product of the number of ways in which we can choose m out of n bins, and the number of ways in which we can put $2m$ balls into m bins—viz.

$$T_1 \ \leq \ \binom{n}{m} \ m^{2m} \ \leq \ n^m \, m^{2m}.$$

Turning now to the second term T_2, note the following inequality: for any numbers $(v_1, \ldots, v_\ell) \in \mathbb{R}_+^\ell$ and non-negative integers $(a_1, \ldots, a_\ell)$, we have

$$\prod_{k=1}^\ell v_k^{a_k} \ \leq \ (\max_{k=1,\ldots,\ell} v_k)^{a_+} \ \leq \ \sum_{k=1}^\ell v_k^{a_+}, \qquad \text{where } a_+ := \sum_{k=1}^\ell a_k.$$

Using this inequality, for any $a \in \mathcal{A}_{-1}$, we have

$$\begin{aligned}\prod_{k=1}^{n} \mathbb{E}\big[(W_{ij}^{(k)})^{a_k}\big] &\leq \prod_{k=1}^{n} \mathbb{E}\big[|W_{ij}^{(k)}|^{a_k}\big] \\ &\leq \sum_{\{k \,|\, a_k \neq 0\}} \mathbb{E}\big[(W_{ij}^{(k)})^{2m}\big] \\ &\leq m \max_{k \in \{1,\ldots,n\}} \mathbb{E}\big[(W_{ij}^{(k)})^{2m}\big],\end{aligned}$$

where the last inequality follows since any multi-index $a \in \mathcal{A}_{-1}$ has at most m non-zero entries. Thus, we have shown that $T_2 \leq m \max_{k \in \{1,\ldots,n\}} \mathbb{E}\big[(W_{ij}^{(k)})^{2m}\big]$.

Substituting our bounds on T_1 and T_2 into equation (85), we obtain

$$\mathbb{E}\big[\big|\sum_{k=1}^{n} \big(W_{ij}^{(k)}\big)^{2m}\big|\big] \leq n^m \, m^{2m+1} \max_{k \in \{1,\ldots,n\}} \mathbb{E}\big[(W_{ij}^{(k)})^{2m}\big]. \tag{86}$$

It thus remains to bound the moments of $W_{ij}^{(k)}$. We have

$$\mathbb{E}\big[(W_{ij}^{(k)})^{2m}\big] \leq \mathbb{E}\big[(X_i^{(k)} X_j^{(k)} - \Sigma_{ij}^*)^{2m}\big] \leq 2^{2m}\big\{\mathbb{E}\big[\big(X_i^{(k)} X_j^{(k)}\big)^{2m}\big] + [\Sigma_{ij}^*]^{2m}\big\},$$

where we have used the inequality $(a+b)^{2m} \leq 2^{2m}(a^{2m} + b^{2m})$, valid for all real numbers a and b. An application of the Cauchy-Schwarz inequality yields

$$\begin{aligned}\mathbb{E}\big[(W_{ij}^{(k)})^{2m}\big] &\leq 2^{2m}\big\{\sqrt{\mathbb{E}\big[\big(X_i^{(k)}\big)^{4m}\big]\,\mathbb{E}\big[\big(X_j^{(k)}\big)^{4m}\big]} + [\Sigma_{ij}^*]^{2m}\big\} \\ &\leq 2^{2m}\big(K_m[\Sigma_{ij}^*\Sigma_{ij}^*]^m + [\Sigma_{oj}^*]^{2m}\big),\end{aligned}$$

where we have used the assumed moment bound $\mathbb{E}\big[\big(X_i^{(k)}/\sqrt{\Sigma_{ii}^*}\big)^{4m}\big] \leq K_m$. Equation (86) thus reduces to

$$\mathbb{E}\big[\big|\sum_{k=1}^{n} W_{ij}^{(k)}\big|^{2m}\big] \leq n^m \, m^{2m+1} \, 2^{2m}\big(K_m[\Sigma_{ii}^*\Sigma_{jj}^*]^m + [\Sigma_{ij}^*]^{2m}\big).$$

Substituting back into equation (84) yields

$$\begin{aligned}\mathbb{P}\Big[\Big|\frac{1}{n}\sum_{k=1}^{n}\Big(X_i^{(k)} X_j^{(k)} - \Sigma_{ij}^*\Big)\Big| > \nu\Big] &\leq \frac{m^{2m+1} n^m 2^{2m}\big(K_m[\Sigma_{ii}^*\Sigma_{jj}^*]^m + [\Sigma_{ij}^*]^{2m}\big)}{n^{2m}\,\nu^{2m}} \\ &= \frac{\big\{m^{2m+1} 2^{2m}\,\big(K_m[\Sigma_{ii}^*\Sigma_{jj}^*]^m + [\Sigma_{ij}^*]^{2m}\big)\big\}}{n^m\,\nu^{2m}}.\end{aligned}$$

Noting that Σ_{ij}^*, Σ_{jj}^*, and Σ_{ij}^* are all bounded above by $\max_i \Sigma_{ii}^*$, we obtain

$$\mathbb{P}\Big[\Big|\frac{1}{n}\sum_{k=1}^{n}\Big(X_i^{(k)} X_j^{(k)} - \Sigma_{ij}^*\Big)\Big| > \nu\Big] \leq \frac{\big\{m^{2m+1} 2^{2m} (\max_i \Sigma_{ii}^*)^{2m}\,(K_m + 1)\big\}}{n^m\,\nu^{2m}},$$

as claimed.

References

[1] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 2008.

[2] P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[3] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[4] L. M. Bregman. The relaxation method for finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7: 191–204, 1967.

[5] L.D. Brown. *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[6] V. V. Buldygin and Y. V. Kozachenko. *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI, 2000.

[7] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Numerical Mathematics and Scientific Computation. Oxford University Press, 1988.

[8] A. d'Asprémont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, 30(1):56–66, 2008.

[9] N. El Karoui. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.*, To appear, 2008.

[10] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostat.*, 9(3):432–441, 2007.

[11] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *J. Multivar. Anal.*, 98(2):227–255, 2007.

[12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[13] J.Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[14] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29 (2):295–327, 2001.

[15] I. M. Johnstone and A. Y. Lu. Sparse principal components analysis. *Unpublished Manuscript*, 2004.

[16] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88:365411, 2003.

[17] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[18] N. Meinshausen. A note on the Lasso for graphical Gaussian model selection. *Statistics and Probability Letters*, 78(7):880–884, 2008.

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

[20] J. M. Ortega and W. G. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, NY, 1970.

[21] A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515, 2008.

[22] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Info. Theory*, 51(3):1030–1051, 2006.

[23] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using the Lasso. Technical Report 709, UC Berkeley, May 2006. To appear in IEEE Trans. Info. Theory.

[24] W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

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

[26] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[27] S. Zhou, J. Lafferty, and L. Wasserman. Time-varying undirected graphs. In *21st Annual Conference on Learning Theory (COLT)*, Helsinki, Finland, July 2008.