Berry-Esseen bounds for estimating undirected graphs[*]

Larry Wasserman[†]

Department of Statistics, Carnegie Mellon University
e-mail: larry@stat.cmu.edu

Mladen Kolar

Booth School of Business, University of Chicago
e-mail: mkolar@gmail.com

and

Alessandro Rinaldo

Department of Statistics, Carnegie Mellon University
e-mail: arinaldo@stat.cmu.edu

Abstract: We consider the problem of providing nonparametric confidence guarantees — with finite sample Berry-Esseen bounds — for undirected graphs under weak assumptions. We do not assume sparsity or incoherence. We allow the dimension D to increase with the sample size n. First, we prove lower bounds that show that if we want accurate inferences with weak assumptions then D must be less than n. In that case, we show that methods based on Normal approximations and on the bootstrap lead to valid inferences and we provide new Berry-Esseen bounds on the accuracy of the Normal approximation and the bootstrap. When the dimension is large relative to sample size, accurate inferences for graphs under weak assumptions are not possible. Instead we propose to estimate something less demanding than the entire partial correlation graph. In particular, we consider: cluster graphs, restricted partial correlation graphs and correlation graphs.

MSC 2010 subject classifications: Primary 62H12; secondary 62H10.
Keywords and phrases: Graphical models, high dimensional inference.

1. Introduction

There are many methods for estimating undirected graphs, such as the glasso (Yuan and Lin, 2007; Friedman and Tibshirani, 2007) and sparse parallel regression (Meinshausen and Bühlmann, 2006). While these methods are very useful, they rely on strong assumptions, in particular, sparsity and some type of incoherence assumption. Most methods do not come with any confidence guarantees.

Recently, some papers — such as Liu (2013) and Ren et al. (2013) — have provided confidence guarantees but they still rely on eigenvalue conditions and sparsity. Ravikumar et al. (2011) consider graph recovery with less restrictive tail conditions on the distribution but they still invoke sparsity and incoherence assumptions. A greedy method for discrete random variables is given in Jalali, Johnson and Ravikumar (2011). Again, an incoherence-like assumption (restricted convexity and smoothness) is invoked.

The purpose of this paper is to construct a nonparametric estimator $\widehat{G}$ of an undirected graph G with confidence guarantees that does not make sparsity and incoherence assumptions. Furthermore, we provide Berry-Esseen type bounds on the coverage of the confidence intervals.

The confidence guarantee we seek is

$$P^n(\widehat{G} \subset G) \geq 1 - \alpha - O(r_n) \tag{1}$$

where n is the sample size, P^n denotes the distribution for n observations drawn from P and r_n is an explicit rate. The notation $\widehat{G} \subset G$ means that the edges of $\widehat{G}$ are a subset of the edges of G. This means that, with high probability, there are no false edges. One could use other error measures such as false discovery rates, but we shall use the guarantee given by (1). Of course, setting $\widehat{G}$ to be the empty graph would trivially satisfy (1). So we also want to ensure that the estimator has non-trivial power for detecting edges.

We focus at first on partial correlation graphs: a missing edge means that the corresponding partial correlation is 0. We distinguish two cases. In the first case, the dimension D can increase with n but is smaller than n. In that case we show that Gaussian asymptotic methods and bootstrap methods yield accurate confidence intervals for the partial correlations which then yield confidence guarantees for the graph. The novelty here is that we provide finite sample bounds on the coverage accuracy. (We also show that, in principle, one can construct finite sample intervals, but these intervals turn out to be too conservative to be useful.)

In the second case, D can be large, even larger than n. In this case it is not possible to get valid inferences for the whole graph under weak assumptions. We investigate several ways to handle this case including: cluster graphs, restricted partial correlation graphs and correlation graphs. Again we provide Berry-Essen bounds for the methods.

Contributions. Here is a summary of our contributions:

1. We develop new Berry-Esseen bounds for the delta method and the bootstrap with increasing dimension.
2. The methods provides confidence guarantees.
3. The methods do not depend on Normality or other parametric assumptions.
4. The methods do not require sparsity or incoherence conditions.
5. The methods have valid coverage when the dimension increases with the sample size.
6. The methods are very simple and do not require any optimization.

Related Work. Our approach is similar to the methods in Liu (2013) and Ren et al. (2013). They use tests on partial correlations to estimate an undirected graph. This approach has two advantages over other methods: it eliminates the need to choose a tuning parameter (as in penalized methods like the glasso) and it provides error control for the estimated graph. However, the results in that paper assume conditions like those in most papers on the penalized methods like the lasso, namely, sparsity. These conditions might be reasonable in some situations, but our goal is to estimate the graph without invoking these assumptions. In the special case of fixed dimension, our method is similar to that in Drton and Perlman (2004).

Schäfer et al. (2005), building on work by Ledoit and Wolf (2004), consider a shrinkage approach to estimating graphs. They make no sparsity or incoherence assumptions. Their examples suggest that their approach can work well in high dimensions. Their method introduces a bias-validity tradeoff: large shrinkage biases the partial correlations but have valid asymptotics in high dimensions. Low shrinkage has low bias but compromises the validity of the asymptotics in high dimensions. Shrinkage graphs are beyond the scope of this paper, however.

Previous research for increasing but moderate dimensions includes Portnoy (1988) and Mammen (1993). Our results are very much in the spirit of those papers. However, our emphasis is on finite sample Berry-Esseen style bounds.

Outline. We start with some notation in Section 2. We discuss various assumptions in Section 3. We then establish lower bounds in Section 4. Finite sample methods are presented in Section 5. However, these do not work well in practice. Asymptotic methods for the moderate dimensional case are considered in Section 6. Specifically, we develop a delta method and a bootstrap method that accommodate increasing dimension. Recent results on high dimensional random vectors due to Chernozhukov, Chetverikov and Kato (2012, 2013) play an important role in our analysis. Methods for the high-dimensional case are considered in Section 7. In Section 8 we give some numerical experiments and some examples. Concluding remarks are in Section 9.

2. Notation

Let $Y_1, \ldots, Y_n \in \mathbb{R}^D$ be a random sample from a distribution P. Each $Y_i = (Y_i(1), \ldots, Y_i(D))^T$ is a vector of length D. We allow $D \equiv D$ to increase with n. We do not assume that the Y_i's are Gaussian. If A is a matrix, we will sometimes let A_{jk} denote the (j, k) element of that matrix.

Let $\Sigma \equiv \Sigma(P)$ denote the $D \times D$ covariance matrix of Y_i and let $\Omega = \Sigma^{-1}$. Let $\Theta = \{\theta\}_{jk}$ be the matrix of partial correlations:

$$\theta_{jk} = -\frac{\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}}. \tag{2}$$

Let

$$S = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})(Y_i - \overline{Y})^T \tag{3}$$

be the the the sample covariance matrix and let $\widehat{\Theta}_n$ be the matrix of sample partial correlations. Given a matrix of partial correlations Θ let $G \equiv G(P)$ be the undirected graph with D nodes and such that there is an edge between nodes j and k if and only if $\theta_{jk} \neq 0$. Equivalently, there is an edge if and only if $\Omega_{jk} \neq 0$. In Section 7 we consider other types of graphs.

For any matrix A, let vec(A) denote the vector obtained by stacking the columns of A. We define the following quantities:

$$\mu = \mathbb{E}(Y), \qquad \sigma = \text{vec}(\Sigma), \qquad \omega = \text{vec}(\Omega) \qquad (4)$$

$$s = \text{vec}(S), \qquad \delta = \sqrt{n}(s - \sigma), \qquad \Delta = \sqrt{n}(\overline{Y} - \mu). \qquad (5)$$

If A is $m \times n$ then there is a unique permutation matrix K_{mn} – called the commutation matrix – such that

$$K_{mn}\text{vec}(A) = \text{vec}(A^T). \qquad (6)$$

Let J denote a $D \times D$ matrix of ones. For matrices L and U with the same dimensions, we write $L \leq U$ to mean that $L_{jk} \leq U_{jk}$ for all j, k. If A is $m \times n$ and B is $p \times q$ then the Kronecker product $A \otimes B$ is the $mp \times nq$ matrix

$$\begin{bmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{bmatrix}. \qquad (7)$$

The Frobenius norm of A is denotes by $||A||_F = \sqrt{\sum_{j,k} A_{jk}^2}$, the operator norm by $||A|| = \sup_{||x||=1} ||Ax||$ and the max norm by $||A||_{\max} = \max_{j,k} |A_{jk}|$. Let $||A||_1 = \max_j \sum_{i=1}^{D} |A_{ij}|$ and

$$\|A\| = \sum_{jk} |A_{jk}|. \qquad (8)$$

We let Φ denote the cdf of a standard Normal random variable. Recall that a random vector $X \in \mathbb{R}^k$ is *sub-Gaussian* if there exists $\zeta > 0$ such that, for all $t \in \mathbb{R}^k$,

$$\mathbb{E}e^{t^T(X-\mu)} \leq e^{||t||^2 \zeta^2/2} \qquad (9)$$

where $\mu = \mathbb{E}(X)$. The smallest and largest eigenvalues of a matrix A are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. We write $a_n \preceq b_n$ to mean that there is some $c > 0$, not depending on n, such that $a_n \leq cb_n$ for all large n. We often use C to denote a generic positive constant.

3. Assumptions

In this section we discuss the assumptions we make and we also discuss some of the commonly used assumptions that we will not use.

The Assumptions. In the case where $D < n$ we make the following assumptions:

(A1) Y and vec(YY^T) are sub-Gaussian.
(A2) $0 < a \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq A < \infty$.

(A3) $\lambda_{\min}(T) \geq c_0 > 0$ where T is the asymptotic covariance of $\sqrt{n}(s - \sigma)$ and is given in Equation (23). Also assume that $\min_j \gamma_{jj} > 0$ where γ, the asymptotic variances of the sample partial correlations, is given in (31).

(A4) $\max_j \mathbb{E}|V_i(j)|^3 \leq C$ where $V_i = \text{vec}[(Y_i - \mu)(Y_i - \mu)^T] - \sigma$.

When $D > n$, we first perform a dimension reduction and then we assume (A1)–(A4) on the reduced problem. We remark that the sub-Gaussian assumption is stronger than needed and is made for simplicity.

Assumptions Note Made. Now we discuss the assumptions that are commonly made for this problem, but that we will not use.

(B1) *Incoherence.* A typical incoherence condition is

$$||\Gamma_{S^cS}(\Gamma_{SS})^{-1}||_\infty < 1 \tag{10}$$

where $\Gamma = \Sigma \otimes \Sigma$, S is the set of pairs with edges between them and $||\cdot||_\infty$ is the maximum absolute column sum. (There are other versions of this assumption but they are similar in character.)

(B2) *Sparsity.* The typical sparsity assumption is that the maximum degree d of the graph is $o(\sqrt{n})$.

(B3) *Donut.* It is assumed that each partial correlation is either 0 or is strictly larger than $\sqrt{\log D/n}$, thus forbidding a donut around the origin.

Discussion. Assumptions (A1) and (A4) will hold if Y has thin enough tails. In fact, if the random variables are bounded or if we truncate the data, then (A1) and (A4) are guaranteed to hold. Assumptions (A2) and (A3) are, strictly speaking, not needed but without them the confidence intervals could be very large. Both (A2) and (A3) are, in principal, testable, using methods like those we develop in this paper (namely, by constructing bootstrap confidence intervals for the eigenvalues).

In contrast, (B1)–(B3) are quite strong. They may be reasonable in certain specialized cases. However, for routine data-analysis, we regard these assumptions with some skepticism when $D > n$. They serve to guarantee that many high-dimensional methods will work, but seem unrealistic in practice. Moreover, the assumptions are very fragile. The incoherence assumption is especially troubling although Ren et al. (2013) have been able to weaken it. The donut assumption ensures that non-zero partial correlations will be detected with high probability. To the best of our knowledge, (B1)–(B3) are not testable when $D > n$. Nor are we aware of a single paper that has ever provided evidence that they do hold in real applications with the exception of applications in engineering (signal processing) where the design is generated by the user.

Our goal is to develop methods that have confidence guarantees, with Berry-Esseen bounds, and that avoid these assumptions.

4. Lower bounds

Constructing a graph estimator for which (1) holds is easy: simply set $\widehat{G}$ to be identically equal to the empty graph. Then $\widehat{G}$ will never contain false edges.

But to have a useful estimator we also want to have non-trivial power to detect edges; equivalently, we want confidence intervals for the partial correlations to have width that shrinks with increasing sample size. In this section we find lower bounds on the width of any confidence interval for partial correlations. This reveals constraints on the dimension D as a function of the sample size n. Specifically, we show that (without sparsity) one must have $D < n$ to get consistent confidence intervals. This is not surprising, but we could not find explicit minimax lower bounds for estimating partial correlations so we provide them here.

The problem of estimating a partial correlation is intimately related to the problem of estimating regression coefficients. Consider the usual regression model

$$Y = \beta_1 X_1 + \cdots + \beta_D X_D + \epsilon \tag{11}$$

where $\epsilon \sim N(0, \sigma^2)$ and where we take the intercept to be 0 for simplicity. (Normality is assumed only in this section.) Suppose we want a confidence interval for β_1.

We will need assumptions on the covariance matrix Σ for $X = (X_1, \ldots, X_D)$. Again, since we are interested in the weak assumption case, we do not want to impose strong assumptions on Σ. In particular, we do not want to rule out the case where the covariates are highly correlated. We do, however, want Σ to be invertible. Let $\mathcal{S}$ denote all symmetric matrices and let

$$\mathcal{S}(a, A) = \left\{ \Sigma \in \mathcal{S} : \ a \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq A \right\} \tag{12}$$

where $0 < a \leq A < \infty$. To summarize: $Y = \beta^T X + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, and $\Sigma = \text{Cov}(X) \in \mathcal{S}(a, A)$. Let $\mathcal{P}$ be all such distributions.

A set-valued function C_n is a $1 - \alpha$ confidence interval for β_1 if

$$P^n(\beta_1 \in C_n) \geq 1 - \alpha \tag{13}$$

for all $P \in \mathcal{P}$. Let $\mathcal{C}_n$ denote all $1 - \alpha$ confidence intervals. Let

$$W_n = \sup\{x : \ x \in C_n\} - \inf\{x : \ x \in C_n\} \tag{14}$$

be the width of C_n.

Theorem 1. *Assume that $D \leq n$ and that $\alpha < 1/5$. Then, for all large n,*

$$\inf_{C_n \in \mathcal{C}_n} \sup_{P \in \mathcal{P}} \mathbb{E}(W_n^2) \geq 4\alpha^3 (1 - 5\alpha) \frac{1}{n - D + 1}. \tag{15}$$

Proof. Let us write the model in vectorized form:

$$Y = X\beta + \epsilon \tag{16}$$

where $Y = (Y_1, \ldots, Y_n)^T$, X is $n \times D$, $\beta = (\beta_1, \ldots, \beta_D)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$.

The proof has two stages. First we will derive a lower bound that involves conditioning on X. In the second stage, we integrate over the marginal distribution of X to get an unconditional bound on the width W_n. Let $M = N(0, \Sigma)$

with $\lambda_{\min}(\Sigma) \geq a > 0$. Without loss of generality, assume $\Sigma_{11}^{-1} = 1$. Let $p_0(x, y) = p_0(y|x)m(x)$ and $p_1(x, y) = p_1(y|x)m(x)$ where $p_0(y|x)$ and $p_1(y|x)$ will be specified later. Here, $m(x)$ is the density of M. Let P_0 and P_1 be the distributions with densities p_0 and p_1. Let P denote a generic distribution in $\mathcal{P}$. Now,

$$\inf_{C_n \in \mathcal{C}_n} \sup_{P \in \mathcal{P}} \mathbb{E}(W_n^2) \geq \inf_{C_n \in \mathcal{C}_n} \max_{P \in \{P_0, P_1\}} \mathbb{E}(W_n^2)$$

$$= \inf_{C_n \in \mathcal{C}_n} \max_{P \in \{P_0, P_1\}} \int \mathbb{E}(W_n^2 | X = x) dM(x)$$

$$= \inf_{C_n \in \mathcal{C}_n} \max_{j=0,1} \int R_j(x) dM(x)$$

where $R_j(x) = \mathbb{E}_j(W_n^2 | X = x)$. Let

$$A = \Big\{ x : R_0(x) > R_1(x) \Big\}.$$

For any two real numbers r_0, r_1, we have that $\max\{r_0, r_1\} \geq (r_0 + r_1)/2$. Hence,

$$\int R_0(x) dM(x) \vee \int R_1(x) dM(x) \geq \int_A R_0(x) dM(x) \vee \int_{A^c} R_1(x) dM(x)$$

$$= \int_A [R_0(x) \vee R_1(x)] dM(x) \vee \int_{A^c} [R_0(x) \vee R_1(x)] dM(x)$$

$$\geq \frac{1}{2} \left(\int_A [R_0(x) \vee R_1(x)] dM(x) + \int_{A^c} [R_0(x) \vee R_1(x)] dM(x) \right)$$

$$= \frac{1}{2} \int [R_0(x) \vee R_1(x)] dM(x).$$

Hence,

$$\inf_{C_n \in \mathcal{C}_n} \sup_{P \in \mathcal{P}} \mathbb{E}(W_n^2) \geq \inf_{C_n} \frac{1}{2} \int [\mathbb{E}_0(W_n^2 | X = x) \vee \mathbb{E}_1(W_n^2 | X = x)] dM(x)$$

$$\geq \frac{1}{2} \int \inf_{C_n} \max_{P_0, P_1} \mathbb{E}_P(W_n^2 | X = x) dM(x).$$

Now we fix $X = x \in \mathbb{R}^{n \times D}$ and lower bound $\inf_{C_n} \max_{P_0, P_1} \mathbb{E}_P(W_n | X = x)$. Assume that $x^T x$ is invertible. Consider Equation (16) where the matrix X is taken as fixed. Multiplying each term in the equation by $(x^T x)^{-1} x^T$ we can rewrite the equation as

$$Z = \beta + \xi$$

where, given $X = x$, $\xi \sim N(0, (x^T x)^{-1})$.

Let $S = x^T x$, $b > 0$, $\delta^2 = 4\alpha^2 S_{11}^{-1}$, $\beta_0 = (0, b, \ldots, b)$ and $\widetilde{\beta}_0 = (\delta, b, \ldots, b)$ which now defines P_0 and P_1. The (conditional) Kullback-Leibler distance between $p_0(y|x)$ and $p_1(y|x)$ is

$$\frac{1}{2}(\widetilde{\beta}_0 - \beta_0)^T (x^T x)(\widetilde{\beta}_0 - \beta_0) = 2\alpha^2 S_{11}^{-1} S_{11}.$$

Note that, since $D \leq n$, $x^T x$ is invertible with probability one. The conditional total variation distance is thus bounded above by $\mathrm{TV}(x) \equiv \alpha \sqrt{S_{11}^{-1} S_{11}}$. Let $A_0 = \{0 \in C_n\}$ and $A_1 = \{\delta \in C_n\}$. Note that $A_0 \cap A_1$ implies that $W_n^2 \geq \delta^2$. So, given $X = x$,

$$
\begin{aligned}
P_0(W_n^2 \geq \delta^2 | X = x) &\geq P_0(A_0 \cap A_1 | X = x) \\
&= P_0(A_0 | X = x) + P_0(A_1 | X = x) - P_0(A_0 \cup A_1 | X = x) \\
&\geq P_0(A_0 | X = x) + P_0(A_1 | X = x) - 1 \\
&\geq P_0(A_0 | X = x) + P_1(A_1 | X = x) - 1 - \mathrm{TV}(x).
\end{aligned}
$$

So far, we have a bound on W_n which depends on x. Now we will turn this into an unconditional bound by integrating over the marginal distribution of X. Note that $\int \mathrm{TV}(x) dM(x) \leq \alpha \int \sqrt{S_{11}^{-1} S_{11}} dM(X)$. Now $\int \sqrt{S_{11}^{-1} S_{11}} dM(x) \to 1$ as $n \to \infty$. Thus, for large enough n, $\int \mathrm{TV}(x) dM(x) \leq 2\alpha$. Integrating over $dM(x)$ we have

$$
\begin{aligned}
P_0(W_n^2 \geq \delta^2) &\geq P_0(A_0) + P_1(A_1) - 1 - \int \mathrm{TV}(x) dM(x) \\
&\geq [1 - \alpha] + [1 - \alpha] - 1 - 2\alpha = 1 - 4\alpha.
\end{aligned}
$$

Let $E = \{S_{11}^{-1} \geq \frac{\alpha}{n - D + 1}\}$. Then,

$$
\begin{aligned}
P_0(W_n^2 \geq \delta^2) &= P_0\left(W_n^2 \geq 4\alpha^2 S_{11}^{-1}\right) \\
&= P_0\left(W_n^2 \geq 4\alpha^2 S_{11}^{-1}, E\right) + P_0\left(W_n^2 \geq 4\alpha^2 S_{11}^{-1}, E^c\right) \\
&\leq P_0\left(W_n^2 \geq \frac{4\alpha^3}{n - D + 1}\right) + P_0\left(E^c\right).
\end{aligned}
$$

Then,

$$
\begin{aligned}
P_0(E^c) &= P_0\left(S_{11}^{-1} < \frac{\alpha}{n - D + 1}\right) = P_0\left(\frac{1}{S_{11}^{-1}} > \frac{n - D + 1}{\alpha}\right) \\
&= P_0\left(\chi_{n-D+1}^2 > \frac{n - D + 1}{\alpha}\right) \leq \alpha
\end{aligned}
$$

by Markov's inequality. Note that the χ^2 random variable is well-defined since $D \leq n$. So

$$
P_0\left(W_n^2 \geq \frac{4\alpha^3}{n - D + 1}\right) \geq P_0(W_n^2 \geq \delta^2) - \alpha \geq 1 - 4\alpha - \alpha = 1 - 5\alpha.
$$

By Markov's inequality,

$$
E_0(W_n^2) \geq (1 - 5\alpha) \frac{4\alpha^3}{n - D + 1}.
$$

$\square$

Now we establish the analogous upper bound.

Theorem 2. *Assume that $D \leq n$ and that $0 < \alpha < 1$. Then a $1 - \alpha$ confidence interval for β_1 is*

$$\left[\widehat{\beta}_1 - \sqrt{2\sigma^2 \log(4/\alpha) \frac{\Sigma_{jj}^{-1}}{(1+x)(n-D+1)}}, \ \widehat{\beta}_1 + \sqrt{2\sigma^2 \log(4/\alpha) \frac{\Sigma_{jj}^{-1}}{(1-x)(n-D+1)}} \right]$$

where $x = \sqrt{\frac{16}{3} \frac{\log(4\alpha^{-1})}{n-D+1}}$. Hence,

$$\inf_{C_n \in \mathcal{C}_n} \sup_{P \in \mathcal{P}} \mathbb{E}(W_n^2) = O\left(\frac{1}{n-D+1}\right). \tag{17}$$

Proof. The least squares estimator is

$$\widehat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon = \beta + Z.$$

We want a $1 - \alpha$ probability bound on $|\widehat{\beta}_j - \beta_j| = |Z_j|$. Since $Z|X \sim N(0, \sigma^2(X'X)^{-1})$, we have that

$$|Z_j| \leq \sqrt{2\sigma^2 (X^T X)_{jj}^{-1} \log(4\alpha^{-1})}$$

with probability $1 - \alpha/2$, conditional on X. We have that $X^T X \sim \text{Wishart}_D(\Sigma, n)$ and

$$\frac{\Sigma_{jj}^{-1}}{(X'X)_{jj}^{-1}} \sim \chi^2_{n-D+1}.$$

For $T \sim \chi^2_D$, we have

$$P(|D^{-1}T - 1| \geq x) \leq \exp^{-\frac{3}{16} D x^2}.$$

Therefore, setting $x = \sqrt{\frac{16}{3} \frac{\log(4\alpha^{-1})}{n-D+1}}$,

$$\frac{\Sigma_{jj}^{-1}}{(1+x)(n-D+1)} \leq (X'X)_{jj}^{-1} \leq \frac{\Sigma_{jj}^{-1}}{(1-x)(n-D+1)}$$

with probability $1 - \alpha/2$. Combining the results, gives a the bound on $|Z_j|$. $\square$

Now consider estimating a partial correlation corresponding to a covariance matrix Σ.

Theorem 3. *Let $W \in \mathbb{R}^D$ where $W \sim N(0, \Sigma)$ with $\Sigma \in \mathcal{S}(a, S)$ (defined in (12)). Let θ be the partial correlation between two components of W, say, W_D and W_{D-1}. Let $\mathcal{C}_n$ be the set of $1 - \alpha$ confidence intervals for θ. Assume that $D \leq n$ and that $\alpha < 1/5$. Then*

$$\inf_{C_n \in \mathcal{C}_n} \sup_{P \in \mathcal{P}} \mathbb{E}(W_n^2) \geq 4\alpha^3 (1 - 5\alpha) \frac{1}{n-D+1}. \tag{18}$$

Proof. Let $b > 0$ be a small positive constant. Let $W = (W_1, \ldots, W_D)$ where

$$W_1 = \epsilon_1$$
$$W_2 = bW_1 + \epsilon_2$$
$$W_3 = bW_2 + bW_1 + \epsilon_3$$
$$\vdots = \vdots$$
$$W_D = qW_{D-1} + bW_{D-2} + \cdots + bW_1 + \epsilon_D,$$

$\epsilon_1, \ldots, \epsilon_D \sim N(0, 1)$. For P_0 take $q = 0$ and for P_1 take $q = \delta$. So, $P_0 = N(0, \Sigma_0)$ and $P_1 = N(0, \Sigma_1)$, say. Then $\Omega_1 = \Sigma_1^{-1}$ corresponds to a complete graph while $\Omega_0 = \Sigma_0^{-1}$ has a missing edge. Let us write $W = (Y, X)$ where $Y = W_1$ and $X = (W_2, \ldots, W_D)$. We note that the marginal distribution of X is the same under P_0 and P_1. The conditional distribution of Y given X under P_j can be written

$$Y = \beta_j^T X + \epsilon$$

where $\beta_0 = (0, b, \ldots, b)$ and $\beta_1 = (\delta, b, \ldots, b)$. The rest of the proof follows the proof of Theorem 1. $\qquad\square$

We conclude that without further assumptions (namely sparsity plus incoherence) we cannot make reliable inferences unless $D \leq n$.

Remark 4. These lower bounds were computed under the assumption of Normality. This is good enough to show the dependence on dimension. However, this makes the minimax lower bound optimistic. When we develop the methods, we shall not assume Normality.

5. A finite sample method

For completeness, we give here a finite sample confidence interval that has length $O(\sqrt{D/n})$. However, the intervals do not work well in practice and we explore asymptotic methods in the following section. In this section we suppose that $|Y_i(j)| \leq B$ for some finite constant B. First we recall the following result from Vershynin (2010).

Theorem 5 (Vershynin, 2010). *There exists c_α, depending only on B, such that*

$$P\left(||S - \Sigma|| > c_\alpha \sqrt{\frac{D}{n}}\right) \leq \alpha.$$

Theorem 6. *Let*

$$\epsilon_n = \frac{c_\alpha}{\widehat{\lambda}^2} \sqrt{\frac{D}{n}} \left(1 - \frac{c_\alpha}{\widehat{\lambda}} \sqrt{\frac{D}{n}}\right)^{-1} \tag{19}$$

where $\widehat{\lambda}$ is the smallest eigenvalue of S. Let $\Delta_n = 2\epsilon_n/(1 - \epsilon_n)$. Then

$$\inf_{P \in \mathcal{P}} P(\underline{\Theta} \leq \Theta \leq \overline{\Theta}) \geq 1 - \alpha \tag{20}$$

where $\overline{\Theta} = \widehat{\Theta} + \Delta_n J$ and $\underline{\Theta} = \widehat{\Theta} - \Delta_n J$ where we recall that J is a $D \times D$ matrix of ones.

Proof. By the previous result, $||S - \Sigma|| \leq c_\alpha \sqrt{\frac{D}{n}}$ with probability at least $1 - \alpha$. From Horn and Johnson (1990) page 381,

$$||S^{-1} - \Sigma^{-1}||_{\max} \leq \frac{||S^{-1}|| \, ||S^{-1}(\Sigma - S)||}{1 - ||S^{-1}(\Sigma - S)||}.$$

Note that, with probability at least $1 - \alpha$,

$$||S^{-1}(\Sigma - S)|| \leq ||S^{-1}|| \, ||\Sigma - S|| = \frac{||\Sigma - S||}{\widehat{\lambda}} \leq \frac{c_\alpha}{\widehat{\lambda}} \sqrt{\frac{D}{n}}. \tag{21}$$

Also note that $||S^{-1}|| \leq 1/\widehat{\lambda}$. We conclude that

$$||S^{-1} - \Sigma^{-1}||_{\max} \leq \epsilon_n.$$

From Lemma 3 of Harris and Drton (2012), $||\widehat{\Theta} - \Theta||_{\max} \leq \frac{2\delta}{1-\delta}$ where $\delta = ||S^{-1} - \Sigma^{-1}||_{\max}$. The result follows. □

Despite the apparent optimal rate, in practice the confidence intervals are gigantic. Instead, we turn to asymptotic methods.

6. Increasing dimension and Berry-Esseen bounds

We call the case where D is increasing with n but smaller than n, the moderate dimensional case. Here we derive confidence sets for the partial correlations in this case. We deal with the high-dimensional case $D > n$ in the next section.

Our goal is to show the accuracy of the delta method and the bootstrap. In particular, we develop new results on the delta method for multiple non-linear statistics with increasing dimension. The state-of-the-art for delta method results are the papers by Pinelis and Molzon (2013); Chen and Shao (2007) where, in particular, the former applies to the multivariate case. Rather than adapt those results, we instead develop a slightly different approach that leverages recent developments in high dimensional statistics. This allows us to develop a simultaneous delta method and bootstrap for multiple inference with increasing dimension. Throughout this section, we assume that $D < n$.

In Sections 6.2–6.6 we assume that (A1)–(A4) hold.

6.1. Preliminary definitions and results

Recall that $s = \text{vec}(S)$, $\sigma = \text{vec}(\Sigma)$, $\omega = \text{vec}(\Omega)$, $\theta = \text{vec}(\Theta)$ and $\delta = \sqrt{n}(s - \sigma)$. Define the map g_j by $\theta_j = g_j(\sigma)$. We can write $\theta = G(\sigma)$ where $G(\sigma) = (g_1(\sigma), \ldots, g_{D^2}(\sigma))^T$. Note that $G : \mathbb{R}^{D^2} \to \mathbb{R}^{D^2}$.

If D is fixed, the central limit theorem implies that

$$\sqrt{n}(s - \sigma) \rightsquigarrow N(0, T) \tag{22}$$

where

$$T \equiv T(\sigma) = \mathbb{E}(\epsilon \epsilon^T \otimes \epsilon \epsilon^T) - \sigma \sigma^T \tag{23}$$

and $\epsilon \sim N(0, \Sigma)$. The finite sample variance matrix of δ is given by (Boik and Haaland (2006)),

$$T_n(\sigma) = \frac{c_1}{n-1}(\mathbb{E}(\epsilon\epsilon^T \otimes \epsilon\epsilon^T) - \sigma\sigma^T) + \left(1 - \frac{D\left(1 - \frac{1}{n}\right)}{n-1}\right)(I_{D^2} - K_{(D,D)})(\Sigma \otimes \Sigma) \tag{24}$$

where $K_{(D,D)}$ is the commutation matrix defined in (6) and $c_1 = D(1 - \frac{1}{n})$.

Let $\widetilde{S} = n^{-1}\sum_{i=1}^n (Y_i - \mu)(Y_i - \mu)^T$, $\widetilde{s} = \text{vec}(\widetilde{S})$, $Q = (\overline{Y} - \mu)(\overline{Y} - \mu)^T$ and $q = \text{vec}(Q)$. Note that

$$s - \sigma = \widetilde{s} - \sigma - q = \overline{V} - q \tag{25}$$

where $\overline{V} = n^{-1}\sum_i V_i$ and $V_i = \text{vec}((Y_i - \mu)(Y_i - \mu)^T) - \sigma$.

Lemma 7. *For all $\epsilon > 0$ we have the following inequalities:*

$$P(||s - \sigma||_\infty > \epsilon) \le 2D^2 e^{-n\zeta^2\epsilon^2/2}$$
$$P(||s - \sigma|| > \epsilon) \le 2D^2 e^{-n\zeta^2\epsilon^2/(2D^2)}$$
$$\mathbb{E}||\delta||_\infty \le \zeta\sqrt{2\log(2D^2)}$$
$$P(||q||_\infty > \epsilon) \le 4D^2 e^{-n\epsilon\zeta^2/2}$$

where ζ is defined in (9).

Proof. Using the sub-Gaussian property, we have $P(||s - \sigma||_\infty > \epsilon) = P(||\overline{V}||_\infty > \epsilon) \le \sum_j P(|\overline{V}_j| > \epsilon) \le 2\sum_j e^{-n\zeta^2\epsilon^2/2} = 2D^2 e^{-n\zeta^2\epsilon^2/2}$. The second result follows from the first since $||s - \sigma|| \le D||s - \sigma||_\infty$. The third inequality follows from a standard inequality; see Lemma 2.2 of Devroye and Lugosi (2001) for example. For the fourth inequality, note that the absolute value $|q_j|$ of each element of q has the form $|\overline{Y}(s) - \mu(s)|\,|\overline{Y}(t) - \mu(t)|$. So $P(||q||_\infty > \epsilon) \le \sum_j P(|q_j| > \epsilon) \le 4D^2 e^{-n\epsilon\zeta^2/2}$. $\square$

Lemma 8. *Let Φ be the cdf of a standard Gaussian random variable. Let A and B be random variables. Then, for every $\epsilon > 0$,*

$$\sup_z |P(A + B < z) - \Phi(z)| \le \sup_z |P(A < z) - \Phi(z)| + \epsilon + P(|B| > \epsilon).$$

Proof. Let $E = \{|B| < \epsilon\}$. Then

$$
\begin{aligned}
P(A + B < z) - \Phi(z) &= P(A + B < z, E) + P(A + B < z, E^c) - \Phi(z) \\
&\le P(A < z + \epsilon) + P(E^c) - \Phi(z) \\
&\le P(A < z + \epsilon) - \Phi(z + \epsilon) - \Phi(z) + \Phi(z + \epsilon) + P(|B| > \epsilon) \\
&\le P(A < z + \epsilon) - \Phi(z + \epsilon) + \epsilon + P(|B| > \epsilon).
\end{aligned}
$$

Hence,

$$\sup_z[P(A + B < z) - \Phi(z)] \leq \sup_z[P(A < z + \epsilon) - \Phi(z + \epsilon)] + \epsilon + P(|B| > \epsilon)$$
$$= \sup_z[P(A < z) - \Phi(z)] + \epsilon + P(|B| > \epsilon).$$

By a similar argument,

$$\sup_z[P(A + B < z) - \Phi(z)] \geq \sup_z[P(A < z) - \Phi(z)] - \epsilon - P(|B| > \epsilon). \qquad \square$$

We need the following recent results on high-dimensional random vectors due to Chernozhukov, Chetverikov and Kato (2012).

Theorem 9 (High-Dimensional CLT; CCK 2012). *Let $Y_1, \ldots, Y_n \in \mathbb{R}^k$ be random vectors with mean μ and covariance Σ and finite third moments. Let*

$$T = \max_j \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i(j) - \mu(j)) \right|.$$

Let $Z \in \mathbb{R}^D$ be Gaussian with mean 0 and covariance Σ. Then

$$\sup_z \left| P(T \leq z) - P(\max_j |Z_j| \leq z) \right| \preceq M \frac{(\log D)^{7/8}}{n^{1/8}} \tag{26}$$

where $M = (E \max_j [|Y(j)| + |Z(j)|]^3)^{1/4}$. Under the sub-Gaussian assumption, $M \preceq (\log D)^{1/8}$. Hence the upper bound is $\log D/n^{1/8}$.

Theorem 10 (Gaussian Anti-Concentration; CCK 2013). *Let $Z_1, \ldots, Z_k$ be centered, not necessarily independent, Gaussian random variables. Then*

$$\sup_z P\left(|\max_j Z_j - z| \leq \epsilon \right) \leq C\epsilon\sqrt{\log(k/\epsilon)} \tag{27}$$

where C depends only on $\max_j \text{Var}(Z_j)$ and $\min_j \text{Var}(Z_j)$.

An immediate corollary of this result is the following.

Lemma 11. *Let $Z \sim N(0, \Sigma)$ where $Z \in \mathbb{R}^k$. There exists $c > 0$ depending only on $\max_j \Sigma_{jj}$ and $\min_j \Sigma_{jj}$ but not on k such that, for every $\epsilon > 0$,*

$$\sup_t \left[P\left(\max_j |Z_j| \leq t + \epsilon \right) - P\left(\max_j |Z_j| \leq t \right) \right] \leq c\epsilon\sqrt{\log(k/\epsilon)}$$

and

$$\sup_t [P(\max_j Z_j \leq t + \epsilon) - P(\max_j Z_j \leq t)] \leq c\epsilon\sqrt{\log(k/\epsilon)}.$$

Proof. Let $Y = \max_j Z_j$. Then

$$P\left(\max_j Z_j \leq t + \epsilon\right) - P\left(\max_j Z_j \leq t\right) \leq P(t - \epsilon \leq Y \leq t + \epsilon)$$

$$= P(-\epsilon \leq Y - t \leq \epsilon)$$

$$\leq P(|Y - t| \leq \epsilon)$$

$$\leq 2 \sup_z P(|Y - z| \leq \epsilon) \leq c\epsilon\sqrt{\log(k/\epsilon)}$$

where the last inequality is precisely the previous anti-concentration inequality. $\quad\square$

Remark 12. A union bound would have given a bound of order $k\epsilon$ instead of $\epsilon\sqrt{\log k/\epsilon}$. Lemma 11 leads to much sharper bounds in our delta method and bootstrap bounds.

Theorem 13 (Gaussian Comparison; CCK 2013). *Let $X = (X_1, \ldots, X_k) \sim N(0, \Sigma_X)$ and $Y = (Y_1, \ldots, Y_k) \sim N(0, \Sigma_Y)$. Let $\Delta = \max_{j,k} |\Sigma_X(j,k) - \Sigma_Y(j,k)|$. Then*

$$\sup_z \left| P(\max_j X_j \leq z) - P(\max_j Y_j \leq z) \right| \leq C\Delta^{1/3}(1 \vee \log(k/\Delta))^{2/3} \qquad (28)$$

where C is only a function of $\max_j \Sigma_Y(j,j)$ and $\min_j \Sigma_Y(j,j)$.

6.2. Berry-Esseen bounds for high-dimensional delta method

Define

$$B = \left\{ a : \ \|a - \sigma\| \leq C\sqrt{D^2 \log n/n} \right\}. \qquad (29)$$

It follows from Lemma 7 that, for large enough C, $P(s \notin B) \leq 1/n^2$. We assume throughout the analysis that $s \in B$ as the error this incurs is of smaller order than the rest of the error terms. Let Θ and $\widehat{\Theta}$ be the matrix of partial correlations and the matrix of estimated partial correlations. Let $\theta = \text{vec}(\Theta)$ and $\widehat{\theta} = \text{vec}(\widehat{\Theta})$. Recall that

$$\theta = (\theta_1, \ldots, \theta_{D^2})^T = G(\sigma) = (g_1(\sigma), \ldots, g_{D^2}(\sigma)).$$

By Taylor expansion and (25),

$$\sqrt{n}(\widehat{\theta} - \theta) = \sqrt{n}L(s - \sigma) + n^{-1/2}R = \sqrt{n}L\overline{V} - \sqrt{n}Lq + \frac{1}{\sqrt{n}}R \qquad (30)$$

where $L = d\text{vec}(G)/d\sigma^T$ so that L is the $D^2 \times D^2$ matrix whose j^{th} row is $\ell_j \equiv \ell_j(\sigma) \equiv dg_j(\sigma)/d\sigma^T$. Similarly, $R = (R_1, \ldots, R_{D^2})^T$ where $R_j = \frac{1}{2}\delta^T H_j \delta$ and H_j is the Hessian of g_j, evaluated at some point between s and σ. Let

$$\Gamma = \text{Var}(\sqrt{n}L(s - \sigma)) = LT_n L^T \quad \text{and} \quad \gamma = \text{diag}(\Gamma). \qquad (31)$$

Let

$$Z = \sqrt{n}\gamma^{-1/2}(\widehat{\theta} - \theta) = (Z_1, \ldots, Z_{D^2})^T$$

where $Z_j = \sqrt{n}(\widehat{\theta}_j - \theta_j)/e_j$ is the normalized estimate, $e_j = \gamma^{1/2}(j,j) = \sqrt{\ell_j(\sigma)^T T(\sigma)\ell_j(\sigma)}$ and $T(\sigma)$ is defined in (23). The covariance of Z is

$$\widetilde{\Gamma} = \gamma^{-1/2}\Gamma\gamma^{-1/2}.$$

Note that $\widetilde{\Gamma}_{jj} = 1$ for all j.

Theorem 14. *Let $W \sim N(0, \widetilde{\Gamma})$ where $W \in \mathbb{R}^{D^2}$ and let*

$$\gamma_n = \max_j \sup_{a \in B} \frac{\|H_j(a)\|}{\sqrt{\ell_j(a)^T T_n(a)\ell_j(a)}} \quad \text{and} \quad \xi_n = \max_j \sup_{a \in B} ||\gamma^{-1/2}\ell_j(a)||_1.$$

where T_n is defined in (24). Then,

$$\sup_z \left| P(\max_j |Z_j| \leq z) - P(\max_j |W_j| \leq z) \right| \preceq A_n \tag{32}$$

where

$$A_n = \frac{\log D}{n^{1/8}} + \frac{4(\gamma_n + \xi_n)}{\zeta^2}\sqrt{\frac{\log(Dn)}{n}}\sqrt{\log\left(\frac{D\zeta^2}{4(\gamma_n + \xi_n)}\sqrt{\frac{n}{\log(Dn)}}\right)} \tag{33}$$

and we recall that ζ is defined in (9). Hence, if $z_\alpha \equiv -\Phi^{-1}(\alpha/D^2)$ then

$$P(\max_j |Z_j| > z_\alpha) \leq \alpha + A_n.$$

Remark 15. In the above result, the dimension enters mainly through the terms γ_n and ξ_n. Except for these terms, the dependence on D is only logarithmic. We discuss these terms in Section 6.5.

Proof. By (30),

$$Z = \sqrt{n}\gamma^{-1/2}(\widehat{\theta} - \theta) = \sqrt{n}\gamma^{-1/2}L\overline{V} - \sqrt{n}\gamma^{-1/2}Lq + \frac{1}{\sqrt{n}}\gamma^{-1/2}R.$$

Note that $\text{Var}(W_i) = \text{Var}(\sqrt{n}\gamma^{-1/2}L\overline{V})$. Fix $\epsilon > 0$ and let

$$E = \left\{\left|\left|\frac{\gamma^{-1/2}R}{\sqrt{n}}\right|\right|_\infty \leq \epsilon\right\} \quad \text{and} \quad E' = \left\{||\sqrt{n}\gamma^{-1/2}Lq||_\infty \leq \epsilon\right\}.$$

Now

$$P(\max_j |Z_j| \leq z) = P(||\sqrt{n}\gamma^{-1/2}L\overline{V} - \sqrt{n}\gamma^{-1/2}Lq + \frac{1}{\sqrt{n}}\gamma^{-1/2}R||_\infty \leq z)$$

$$\leq P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty - ||\sqrt{n}\gamma^{-1/2}Lq||_\infty$$

$$- ||\frac{1}{\sqrt{n}}\gamma^{-1/2}R||_\infty \leq z)$$

$$= P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty - ||\sqrt{n}\gamma^{-1/2}Lq||_\infty$$

$$- ||\frac{1}{\sqrt{n}}\gamma^{-1/2}R||_\infty \leq z, \, E) + P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty$$

$$- ||\sqrt{n}\gamma^{-1/2}Lq||_\infty - ||\frac{1}{\sqrt{n}}\gamma^{-1/2}R||_\infty \leq z, \, E^c)$$

$$\leq P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty - ||\sqrt{n}\gamma^{-1/2}Lq||_\infty \leq z + \epsilon) + P(E^c)$$

$$= P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty - ||\sqrt{n}\gamma^{-1/2}Lq||_\infty \leq z + \epsilon, E')$$

$$+ P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty - ||\sqrt{n}\gamma^{-1/2}Lq||_\infty \leq z + \epsilon, (E')^c)$$

$$+ P(E^c)$$

$$\leq P(||\sqrt{n}\gamma^{-1/2}L\overline{V}||_\infty \leq z + 2\epsilon) + P(E^c) + P((E')^c).$$

So,

$$P(\max_j |Z_j| \leq z) - P(\max_j |W_j| \leq z)$$

$$\leq P\left(\left|\left|\sqrt{n}\gamma^{-1/2}L\overline{V}\right|\right|_\infty \leq z + 2\epsilon\right) - P(\max_j |W_j| \leq z + 2\epsilon)$$

$$+ P(\max_j |W_j| \leq z + 2\epsilon) - P(\max_j |W_j| \leq z) + P(E^c) + P((E')^c)$$

$$\leq P\left(\left|\left|\sqrt{n}\gamma^{-1/2}L\overline{V}\right|\right|_\infty \leq z + 2\epsilon\right) - P(\max_j |W_j| \leq z + 2\epsilon)$$

$$+ C\epsilon\sqrt{\log D/\epsilon} + P(E^c) + P((E')^c)$$

$$\leq C\frac{\log D}{n^{1/8}} + C\epsilon\sqrt{\log D/\epsilon} + P(E^c) + P((E')^c)$$

where we used Theorem 9 applied to $\overline{V}^* = \gamma^{-1/2}L\overline{V}$ and Lemma 11. Recall that $s \in B$ except on a set of probability $1/n^2$ and on this set,

$$\left(\frac{\gamma^{-1/2}R}{\sqrt{n}}\right)_j = \frac{\delta^T H_j \delta}{\sqrt{n\ell_j^T T_n \ell_j}} \leq \gamma_n \sqrt{n}||s - \sigma||_\infty^2$$

and so by Lemma 7,

$$P(E^c) \leq 2D^2 \exp\left(-\frac{\sqrt{n}\zeta^2\epsilon^2}{2\gamma_n^2}\right).$$

Choosing

$$\epsilon = \frac{4(\gamma_n + \xi_n)}{\zeta^2}\sqrt{\frac{\log(Dn)}{n}}$$

we have $P(E^c) \le \frac{1}{n^2}$ and

$$\epsilon \sqrt{\log D/\epsilon} \le \frac{4(\gamma_n + \xi_n)}{\zeta^2} \sqrt{\frac{\log(Dn)}{n}} \sqrt{\log\left(\frac{D\zeta^2}{4(\gamma_n + \xi_n)} \sqrt{\frac{n}{\log(Dn)}}\right)}.$$

Using Holder's inequality,

$$|\gamma^{-1/2}\ell_j^T q| \le ||q||_\infty \, ||\gamma^{-1/2}\ell_j||_1 \le ||q||_\infty \, \xi_n$$

so that $||\gamma^{-1/2}Lq||_\infty \le ||q||_\infty \, \xi_n$. Hence, using Lemma (7),

$$P((E')^c) \le P(||q||_\infty > \epsilon/\sqrt{\xi_n n}) \le 4D^2 e^{-\sqrt{n}\zeta^2\epsilon/(2\xi_n)} \le \frac{1}{n^2}$$

The result follows by computing a similar lower bound and taking the supremum over z. For the last statement, note that $W_j \sim N(0,1)$. So

$$P(\max_j |Z_j| > z_\alpha) \le P(\max_j |W_j| > z_\alpha) + A_n \le \sum_j P(|W_j| > z_\alpha) + A_n \le \alpha + A_n.$$

$\square$

In practice we need to use $T_j = \sqrt{n}(\widehat{\theta}_j - \theta_j)/\widehat{e}_j$ where $\widehat{e}_j = \sqrt{\ell_j(s)^T T(s)\ell_j(s)} \equiv U_j(s)$ is the estimated standard error. We have the following result for this case.

Theorem 16. *Define γ_n and ξ_n as in the previous theorem. Let*

$$\rho_n = \max_j \sup_{a \in B} \frac{||U_j'(a)||_1}{\sqrt{\ell_j(a)^T T_n(a)\ell_j(a)}}$$

where $U_j(a) = \sqrt{\ell_j^T(a)T(a)\ell_j(a)}$. Then,

$$\sup_z |P(\max_j |T_j| \le z) - P(\max_j |W_j| \le z)| \preceq A_n + \rho_n \sqrt{\frac{\log n}{n}}$$

where A_n is defined in (33). If $z \equiv -\Phi^{-1}(\alpha/D^2)$ then

$$\sup_z |P(\max_j |T_j| > z)| \le \alpha + A_n + \rho_n \sqrt{\frac{\log n}{n}}.$$

Proof. Let $E = \{\max_j e_j/\widehat{e}_j < 1 + \epsilon\}$ and $F = \{\max Z_j < u/\epsilon\}$ where $\epsilon = (4\rho_n/\zeta)\sqrt{\log n/(n\zeta^2)}$ and $u = \epsilon\sqrt{\log(n)}$. Note that $e_j - \widehat{e}_j = U_j(\sigma) - U_j(s) = (\sigma - s)^T U_j'$ where U' is the gradient of U evaluated at some point between s and σ. Then, for $0 < \epsilon \le 1$,

$$P(E^c) \le P\left(\max_j \frac{e_j - \widehat{e}_j}{e_j} > \frac{\epsilon}{1+\epsilon}\right) = P\left(\max_j \frac{U_j(\sigma) - U_j(s)}{e_j} > \frac{\epsilon}{1+\epsilon}\right)$$

$$= P\left(\max_j \frac{(\sigma - s)^T U_j'}{e_j} > \frac{\epsilon}{1+\epsilon}\right) \le P\left(\frac{||s - \sigma||_\infty \max_j ||U_j'||_1}{e_j} > \frac{\epsilon}{1+\epsilon}\right)$$

$$\leq P\left(||s-\sigma||_\infty \rho_n > \frac{\epsilon}{1+\epsilon}\right) = P\left(||s-\sigma||_\infty > \frac{\epsilon}{2\rho_n}\right)$$

$$\leq D^2 e^{-n\epsilon^2/(2\rho_n^2)} \leq \frac{1}{n^2}.$$

Now,

$$P\left(\max_j \frac{\sqrt{n}(\widehat{\theta}_j - \theta_j)}{\widehat{e}_j} \leq z\right) - P(\max W_j \leq z)$$

$$= P\left(\max_j Z_j\left(\frac{e_j}{\widehat{e}_j}\right) \leq z\right) - P(\max W_j \leq z)$$

$$\leq P\left(\max_j Z_j(1-\epsilon) \leq z\right) + P(E^c) - P(\max W_j \leq z)$$

$$= P\left(\max_j Z_j - Z_j\epsilon \leq z\right) + P(E^c) - P(\max W_j \leq z)$$

$$\leq P\left(\max_j Z_j \leq z + u\right) + P(F^c) + P(E^c) - P(\max W_j \leq z)$$

$$\leq P\left(\max_j Z_j \leq z + u\right) - P(\max W_j \leq z + u)$$
$$+ Cu\sqrt{\log D/u} + P(F^c) + P(E^c)$$

$$\leq \sup_z \left[P\left(\max_j Z_j \leq z\right) - P(\max_j W_j \leq z)\right]$$
$$+ Cu\sqrt{\log D/u} + P(F^c) + P(E^c)$$

$$\leq A_n + Cu\sqrt{\log D/u} + P(F^c) + P(E^c)$$

where A_n is defined in (33). Next,

$$P(F^c) = P(\max_j Z_j > u/\epsilon) \leq P(\max_j W_j > u/\epsilon) + A_n$$

$$= P(\max_j W_j > \sqrt{\log n}) + A_n$$

$$\leq \frac{E(\max_j W_j)}{\sqrt{\log n}} + A_n \preceq \frac{\sqrt{\log D}}{\sqrt{\log n}} + A_n \preceq A_n.$$

So

$$\sup_z |P(\max_j |T_j| \leq z) - P(\max_j |W_j| \leq z)|$$

$$\preceq \sup_z [P\left(\max_j Z_j \leq z\right) - P(\max W_j \leq z)] + A_n$$
$$+ \frac{1}{n^2} + Cu\sqrt{\log D}$$
$$\preceq A_n + \rho_n\sqrt{\frac{\log n}{n}}.$$

A similar lower bound completes the proof. $\square$

6.3. The bootstrap

In this section we assume that $\max_j |Y(j)| \leq B$ for some $B < \infty$. This is not necessary but it simplifies the proofs. We do not require that B be known. Let $Y_1^*, \ldots, Y_n^*$ be a sample from the empirical distribution and let s^* be the corresponding (vectorized) sample covariance. Now let $\widehat{\theta}^*$ be the partial correlations computed from $Y_1^*, \ldots, Y_n^* \sim P_n$ where P_n is the empirical distribution. The (un-normalized) bootstrap rectangle for θ is

$$\mathcal{R}_n = \left\{ \theta : ||\theta - \widehat{\theta}||_\infty \leq \frac{Z_\alpha}{\sqrt{n}} \right\}$$

where $Z_\alpha = \widehat{F}^{-1}(1 - \alpha)$ and

$$\widehat{F}(z) = P\left(\sqrt{n}||\widehat{\theta}^* - \widehat{\theta}||_\infty \leq z \mid Y_1, \ldots, Y_n \right) \tag{34}$$

is the bootstrap approximation to

$$F(z) = P(\sqrt{n}||\widehat{\theta} - \theta||_\infty \leq z).$$

The accuracy of the coverage of the bootstrap rectangle depends on $\sup_z |\widehat{F}(z) - F(z)|$.

Let

$$\Gamma = \text{Var}(\sqrt{n}L(s - \sigma)) = LT_nL^T.$$

Let $Z \sim N(0, \Gamma)$ where $Z \in \mathbb{R}^{D^2}$. First we need the following limit theorem for the un-normalized statistics.

Theorem 17. *Define $\gamma_n' = \max_j \sup_{a \in B} \|H_j(a)\|$ and $\xi_n' = \max_j \sup_{a \in B} ||\ell_j(a)||_1$. Then*

$$\sup_z \left| P(\sqrt{n}||\widehat{\theta} - \theta||_\infty \leq z) - P(||Z||_\infty \leq z) \right| \preceq \frac{\log D}{n^{1/8}} + A_n'$$

where

$$A_n' = \frac{\log D}{n^{1/8}} + \frac{4(\gamma_n' + \xi_n')}{\zeta^2} \sqrt{\frac{\log(Dn)}{n}} \sqrt{\log\left(\frac{D\zeta^2}{4(\gamma_n' + \xi_n')} \sqrt{\frac{n}{\log(Dn)}} \right)}. \tag{35}$$

Proof. The proof is the same as the proof of Theorem 14 with γ_n' and ξ_n' replacing γ_n and ξ_n. $\qquad\qquad\square$

Now we bound $\sup_z |\widehat{F}(z) - F(z)|$.

Theorem 18.

$$\sup_z |\widehat{F}(z) - F(z)| \preceq \frac{\log D}{n^{1/8}} + (\gamma_n' + \xi_n')\sqrt{\log n/n} + O_P\left(\left(\frac{\log D}{n} \right)^{1/6} \right)$$

and hence

$$P(\theta \notin \mathcal{R}_n) \leq \alpha + \frac{\log D}{n^{1/8}} + (\gamma_n' + \xi_n')\sqrt{\log n/n} + O\left(\left(\frac{\log D}{n}\right)^{1/6}\right).$$

Proof. Let $Z \sim N(0, \Gamma)$ and let $Z' \sim N(0, \Gamma_n)$ where $\Gamma_n = \text{Var}(\sqrt{n}L(s^* - s)|Y_1, \ldots, Y_n)$. Then

$$\sup_z |\widehat{F}(z) - F(z)| \leq \sup_z \left|F(z) - P(||Z||_\infty \leq z)\right| + \sup_z \left|\widehat{F}(z) - P(||Z'||_\infty \leq z)\right|$$
$$+ \sup_z \left|P(||Z'||_\infty \leq z) - P(||Z||_\infty \leq z)\right|$$
$$= I \quad + \quad II \quad + \quad III.$$

In the previous theorem, we showed that $I \leq \frac{\log D}{n^{1/8}} + A_n'$. For II, we proceed exactly as in the proof for of the previous theorem but with P_n replacing P (and with $Y_1, \ldots, Y_n$ fixed). This yields, for any $\epsilon > 0$,

$$\widehat{F}(z) - P(||Z'||_\infty \leq z) \preceq \frac{\log D}{n^{1/8}} + \epsilon\sqrt{\log D/\epsilon}$$
$$+ \quad P(\sqrt{n}||Lq^*||_\infty > \epsilon|Y_1, \ldots, Y_n)$$
$$+ \quad P(n^{-1/2}||R^*||_\infty > \epsilon|Y_1, \ldots, Y_n)$$

where $q^* = \text{vec}((\overline{Y}^* - \overline{Y})(\overline{Y}^* - \overline{Y})^T)$, $R_j^* = (1/2)\delta^T H_j^* \delta^*$, $\delta^* = \sqrt{n}(s^* - s)$ and H_j^* is the Hessian of g_j evaluated at a point between s and s^*.

Since all the Y_i's are contained in the bounded rectangle $B \times \cdots \times B$, it follows that under the empirical measure P_n, Y_i^* is sub-Gaussian with $\zeta = B$. It then follows that $s^* \in B$ expect on a set of probability at most $1/n$. Choosing

$$\epsilon = \frac{4(\gamma_n' + \xi_n')}{B^2}\sqrt{\frac{\log(Dn)}{n}}$$

and arguing as in the proof of Theorem 14 we conclude that

$$\widehat{F}(z) - P(||Z'||_\infty \leq z) \preceq \frac{\log D}{n^{1/8}} + \epsilon\sqrt{\log D/\epsilon}$$
$$+ \quad P(\sqrt{n}||Lq^*||_\infty > \epsilon|Y_1, \ldots, Y_n)$$
$$+ \quad P(n^{-1/2}||R^*||_\infty > \epsilon|Y_1, \ldots, Y_n)$$
$$\leq \frac{\log D}{n^{1/8}} + O_P(A_n').$$

For III, we use Theorem 13 which implies that

$$III \leq C \Delta^{1/3}(1 \vee \log(k/\Delta))^{2/3}$$

where $\Delta = \max_{s,t} |\Gamma(s,t) - \Gamma_n(s,t)|$. Each element of $\Gamma_n(s,t)$ is a sample moment and $\Gamma(s,t)$ is corresponding population moment, and so, since P_n is sub-Gaussian, $\Delta = O_P(\sqrt{\log D/n})$. Hence, $III = O_P(\frac{\log D}{n})^{1/6}$. $\qquad\square$

6.4. A super-accurate bootstrap

Now we describe a modified approach to the bootstrap that has coverage error only $O(\log D/n^{1/8})$ which is much more accurate than the usual bootstrap as described in the last section. The idea is very simple. Let $\mathcal{R}$ be the $1 - \alpha$ bootstrap confidence rectangle for σ described in Section 7.1. Write $\theta = G(\sigma)$ and define

$$\mathcal{T} = \Big\{ G(\sigma) : \ \sigma \in \mathcal{R} \Big\}.$$

By construction, $\mathcal{T}$ inherits the coverage properties of $\mathcal{R}$ and so we have immediately:

Corollary 19.

$$P(\theta \in \mathcal{T}) \geq 1 - \alpha - O\left(\frac{\log D}{n^{1/8}}\right) - O\left(\frac{\log D}{n}\right)^{1/6}.$$

The set $\mathcal{T}$ then defines confidence sets for each θ_j, namely,

$$C_j = \Big[\inf\{g_j(\sigma): \ \sigma \in \mathcal{R}\}, \quad \sup\{g_j(\sigma): \ \sigma \in \mathcal{R}\}\Big].$$

We should stress that, in general, obtaining a confidence set by mapping a confidence rectangle can lead to wide intervals. However, our foremost concern in this paper is coverage accuracy.

Constructing the set $\mathcal{T}$ can be difficult. But it is easy to get an approximation. We draw a large sample $\sigma_1, \ldots, \sigma_N$ from a uniform distribution on the rectangle $\mathcal{R}$. Now let

$$\underline{\theta}_j = \min_{1 \leq s \leq N} g_j(\sigma_s), \quad \overline{\theta}_j = \max_{1 \leq s \leq N} g_j(\sigma_s).$$

Then $[\underline{\theta}_j, \overline{\theta}_j]$ approximates the confidence interval for θ_j. Alternatively, we take $\sigma_1, \ldots, \sigma_N$ to be the bootstrap replications that are contained in $\mathcal{R}$. Note that there is no need for a multiple comparison correction as the original confidence rectangle is a simultaneous confidence set.

6.5. Comments on the error terms

The accuracy of the delta method depends on the dimension D mainly through the terms γ_n, ξ_n and ρ_n. Similarly, the accuracy of the (first version of the) bootstrap depends on γ_n' and ξ_n'. In this section we look at the size of these terms. We focus on γ_n' and ξ_n'.

Recall that $\ell_j = d\theta_j/d\sigma^T$. Then

$$\ell_j(\sigma) = \frac{d\theta_j}{d\sigma^T} = \frac{d\theta_j}{d\omega^T} \frac{d\omega}{d\sigma^T}.$$

Let (s, t) be such that $\theta_j = \Theta_{st}$. Then, $\frac{d\theta_j}{d\omega^T}$ is $1 \times D^2$ and $\frac{d\omega}{d\sigma^T}$ is $D^2 \times D^2$. Now $\frac{d\omega}{d\sigma^T} = -\Omega \otimes \Omega$ and $\frac{d\theta_j}{d\omega^T}$ is 0 except for three entries, namely,

$$\frac{d\theta_j}{d\Omega_{ss}} = -\frac{\theta_j}{2\Omega_{ss}}, \qquad \frac{d\theta_j}{d\Omega_{tt}} = -\frac{\theta_j}{2\Omega_{ss}}, \qquad \frac{d\theta_j}{d\Omega_{st}} = \frac{\theta_j}{\Omega_{st}}.$$

Define (J, K, M) by $\sigma_J = \Sigma_{ss}$, $\sigma_K = \Sigma_{tt}$ and $\sigma_M = \Sigma_{st}$. Then

$$\ell_j = \frac{d\theta_j}{d\sigma^T} = \frac{\theta_j}{2\Omega_{ss}}[\Omega \otimes \Omega]_J + \frac{\theta_j}{\Omega_{st}}[\Omega \otimes \Omega]_M + \frac{\theta_j}{2\Omega_{tt}}[\Omega \otimes \Omega]_K = f_j(\Omega \otimes \Omega) \quad (36)$$

where $[A]_j$ denotes the jth row of A and f_j is a sparse vector that is 0 except for three entries.

Now the Hessian is $H_j = (\frac{d\ell_1}{d\sigma^T}, \ldots, \frac{d\ell_{D^2}}{d\sigma^T})^T$ where

$$\frac{d\ell_j}{d\sigma^T} = \frac{d\ell_j}{d\omega^T}\frac{d\omega}{d\sigma^T} = -\frac{d\ell_j}{d\omega^T}(\Omega \otimes \Omega).$$

Now

$$\frac{d\ell_j}{d\omega^T} = \left(\left(\frac{d\omega}{d\sigma^T}\right)^T \otimes I\right)\frac{d}{d\omega^T}\left(\frac{d\theta_j}{d\omega}\right) + \left(I \otimes \frac{d\theta_j}{d\omega^T}\right)\frac{d}{d\omega^T}\frac{d\omega}{d\sigma^T}$$

$$= -(\Omega \otimes \Omega \otimes I)f_j - (I \otimes f_j)\frac{d}{d\omega^T}(\Omega \otimes \Omega)$$

$$= -(\Omega \otimes \Omega \otimes I)f_j$$
$$- (I \otimes f_j)(I_D \otimes K_{(D,D)} \otimes I_D)(I_{D^2} \otimes \text{vec}(\Omega) : \text{vec}(\Omega) \otimes I_{D^2});$$

where we used the fact that

$$\frac{d\text{vec}(\Omega \otimes \Omega)}{d\omega^T} = (I_D \otimes K_{(D,D)} \otimes I_D)(I_{D^2} \otimes \text{vec}(\Omega) : \text{vec}(\Omega) \otimes I_{D^2});$$

see, for example, p 185 of Magnus and Neudecker (1988). Note that $||f_j||_0 = O(1)$ independent of D. The presence of this sparse vector helps to prevent the gradient and Hessian from getting too large.

By direct examination of ℓ_j and H_j we see that the size of γ'_n and ξ'_n depends on how dense Ω is. In particular, when Ω is diagonally dominant, γ'_n and ξ'_n are both $O(1)$. In this case the error terms have size $O((\log D)/n^{1/8})$. However, if Ω is dense, then $||\ell_j||_1$ can be of order $O(D^2)$ and and $\|H_j\|$ can be of order $O(D^4)$. In this case the error can be as large as $D^4/n^{1/8}$. On the other hand, the bootstrap in Section 6.4 always has accuracy $O((\log D)/n^{1/8})$. But the length of the intervals could be large when Ω is dense. And note that even in the favorable case, we still require $D < n$ for the results to hold. (We conjecture that this can be relaxed by using shrinkage methods as in Schäfer et al. (2005).) These observations motivate the methods in the next section which avoid direct inferences about the partial correlation graph in the high-dimensional case.

It is interesting to compare the size of the errors to other work on inference with increasing dimension. For example, Portnoy (1988) gets accuracy $\sqrt{D^{3/2}/n}$ for maximum likelihood estimators in exponential families and Mammen (1993) gets accuracy $\sqrt{D^2/n}$ for the bootstrap for linear models.

6.6. Back to graphs

Finally, we can use the above methods for estimating a graph with confidence guarantees. We put an edge between j and k only if 0 is excluded from the confidence interval for θ_{jk}. The desired guarantee stated in (1) then holds.

7. The high dimensional case

Now we consider the case where $D > n$. We present three methods for dealing with the high-dimensional case:

(B1) **Correlation graphs**. This is a common technique in biostatistics. We connect two nodes if the confidence interval for two variables excludes $[-\epsilon, \epsilon]$ for some threshold $\epsilon \in [0, 1]$. Our contribution here is to provide confidence guarantees using the bootstrap that are valid as long as $D = o(e^{n^{1/7}})$. In this paper we use $\epsilon = 0$.

(B2) **Cluster graphs**. We cluster the features and average the features within each cluster. These averaged features define a (dimension-reduced) set of new features. As long as the number of clusters L is $o(n)$ we get valid inferences. In this case, each node corresponds to the average of the variables in a cluster and the edges correspond to conditional independence statements about the new features.

(B3) **Restricted Graphs**. Define the restricted partial correlation

$$\theta_{jk} \equiv \sup_{|S| \leq L} |\theta(Y(j), Y(k)|Y(S))|$$

where L is some fixed number, $\theta(Y(j), Y(k)|Y(S))$ is the partial correlation between $Y(j)$ and $Y(k)$ given the set of variables $Y(S)$ where S varies over all subsets of $\{1, \ldots, D\} - \{j, k\}$ of size L These are sometimes called lower-order partial correlations. Now construct a graph based on the restricted partial correlations. Note that $L = 0$ is a correlation graph and $L = D$ is a partial correlation graph. (This is similar to the idea in Castelo and Roverato, 2006). The bootstrap leads to valid inferences only requiring $D = o(e^{n^{1/7}})$.

Remark 20. Following Schäfer et al. (2005), we could estimate $U = (1-\lambda)\Sigma + \lambda T$ where T is, for example, a diagonal matrix. The graph is constructed from biased partial correlations corresponding to U^{-1}. When λ is close to 1, high-dimensional asymptotic confidence intervals have accurate coverage. Thus we have a bias-validity tradeoff. Investigating this tradeoff is quite involved and so we will examine this method elsewhere.

In this section we make the following assumptions.

(A1) Y and $\text{vec}(YY^T)$ are sub-Gaussian.

(A2) $\max_j \mathbb{E}|V_i(j)|^3 \leq C$ where $V_i = \text{vec}[(Y_i - \mu)(Y_i - \mu)^T] - \sigma$.

(A3) $D = o(e^{n^{1/7}})$.

The proofs of the results in this section are similar to those in Section 6 but they are easier as the error terms are, by design, not dependent on dimension sensitive quantities like γ_n and ξ_n. Because of this, we shall only present proof outlines.

(B1) Select a threshold ϵ.

(B2) Compute the sample covariance matrix R.

(B3) Construct a $1 - \alpha$ bootstrap confidence rectangle $\mathcal{R}$ for the correlations.

(B4) Put an edge between nodes j and k if $[-\epsilon, \epsilon]$ is not in the confidence interval for ρ_{jk}.

FIG 1. *The Correlation Graph Algorithm.*

7.1. Correlation graphs

The simplest approach to constructing graphs is to use correlation or covariances rather than partial correlation. Let ρ_{jk} denoted the correlation between $Y(j)$ and $Y(k)$. The true graph G_ϵ connects j and k if $|\rho(j,k)| > \epsilon$ where $0 \leq \epsilon \leq 1$ is some user-specified threshold. The algorithm is in Figure 1. Of course, we can use either ρ or σ; we get the same graph from either.

Theorem 21. *Let r_{jk} denote the sample correlation between $Y(j)$ and $Y(k)$ and let r be the $D^2 \times 1$ vector of correlations. Similarly, let ρ be the vector of true correlations. Define Z_α by the bootstrap equation*

$$P\left(\max_{jk} \sqrt{n}|r_{jk}^* - r_{jk}| > Z_\alpha \;\middle|\; Y_1, \ldots, Y_n\right) = \alpha. \tag{37}$$

Let

$$\mathcal{R} = \left\{a \in \mathbb{R}^{D^2} : \; ||a - r||_\infty \leq \frac{Z_\alpha}{\sqrt{n}}\right\}.$$

Then

$$P(\rho \in \mathcal{R}) \geq 1 - \alpha - O\left(\frac{\log D}{n^{1/8}}\right) - O\left(\frac{\log D}{n}\right)^{1/6}.$$

We thus have

$$P(\widehat{G}_\epsilon \subset G_\epsilon \;\; \text{for all } \epsilon) \geq 1 - \alpha + \frac{\log D}{n^{1/8}} + O\left(\frac{\log D}{n}\right)^{1/6}. \tag{38}$$

Remark 22. A very refined Berry-Esseen result for a single correlation was obtained by Pinelis and Molzon (2013).

Proof Outline. The proof is the same as the proof of Theorem 18. However, in this case, it is easy to see that γ'_n and ξ'_n are $O(1)$, independent of the D since the gradient ℓ_j and Hessian H_j is a function only of the bivariate distribution of $(Y(s), Y(t))$ corresponding to the correlation. □

7.2. Cluster graphs

The idea here is to partition the features into clusters, average the features within each cluster and then form the graph for the new derived features. If the clusters are sufficiently few, then valid inference is possible.

There are many clustering methods. Here we consider choosing a set of representative features — or prototypes — using the L-centers algorithm, which we describe below. Then we assign each feature to its nearest center. We average

(B1) Choose $L = o(n)$.

(B2) Randomly split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$.

(B3) Using $\mathcal{D}_1$ select L proto-features:

    (a) Choose a feature j randomly and set $\mathcal{S} = \{j\}$ and $\mathcal{C} = \{1, \ldots, D\} - \mathcal{S}$.

    (b) Repeat until $\mathcal{S}$ has L elements:

        i. For each $j \in \mathcal{C}$ compute the minimum distance $d_j = \min_{i \in \mathcal{S}} d(i, j)$.

        ii. Find $j \in \mathcal{C}$ to maximize d_j. Move j from $\mathcal{C}$ to $\mathcal{S}$.

    (c) For L clusters by assigning each feature to its closest center.

    (d) Average the features within each clusters.

(B4) Using $\mathcal{D}_2$, construct a confidence graph for the L new features using either the delta method or the bootstrap from Section 6.

(B5) (Optional): Construct a correlation graph for the features within each cluster.

FIG 2. *The Cluster Graph Algorithm.*

the features within each cluster and then find the undirected graph of these new L derived features. Let $\widetilde{G}$ be the graph for these new features. We estimate $\widetilde{G}$ using confidence intervals for the partial correlations. Note that the graph $\widetilde{G}$ as well as the estimated graph $\widehat{G}$ are both random.

To ensure the validity of the confidence intervals, we use data spitting. We split the data randomly into two halves. The first half is used for clustering. The confidence intervals are constructed from the second half of the data.

The cluster-graph algorithm is described in Figure 2. It is assumed in the algorithm that the number of features $L = o(n)$ is specified by the user. An improvement is to use a data-driven approach to choosing L. We leave this to future work.

The asymptotic validity of the method follows from the results in Section 6 together with the data-splitting step. Without the data-splitting step, the proofs in Section 6 would not be valid since the feature selection process would introduce a bias. The independence introduced by the splitting thus seems critical. Whether it is possible to eliminate the data-splitting is an open problem. Let us state, without proof, the validity assuming the bootstrap is used. A similar result holds for the delta method.

Theorem 23. *Let θ be the vector of k partial correlations for the features selected from the first half of the data. Let $\mathcal{R}$ be the confidence rectangle using the second half of the data. Then*

$$P(\theta \notin \mathcal{R}_n) \leq \alpha + \frac{(\log L)}{n^{1/8}} + (\gamma_n' + \xi_n')\sqrt{\log n/n} + O\left(\left(\frac{\log L}{n}\right)^{1/6}\right) \quad (39)$$

where γ_n' and ξ_n' are functions of the distribution of the selected features.

Another possibility is as follows. For each (j, k) let Z_{jk} be a dimension reduction of the variables $(Y(s) : s \neq j, k)$. Then we could estimate the partial

correaltion of $Y(j)$ and $Y(k)$ given Z_{jk}. This would require a separate dimension reduction step for each pair (j, k).

7.3. Restricted partial correlations

Instead of building a graph from partial correlations, we can use a weaker measure of dependence. Motivated by Castelo and Roverato (2006), we define

$$\theta_{jk} = \sup_{|S| \leq L, (j,k) \notin S} |\theta(Y(j), Y(k)|Y(S))|. \tag{40}$$

For $L = 0$ we get a correlation graph. For $L = D-2$ we get back the usual partial correlation graph. By choosing $0 < L = o(n)$ we get something in between these two cases while still retaining validity of the confidence intervals.

The estimate of θ_{jk} is the sample version

$$\widehat{\theta}_{jk} = \sup_{|S| \leq k} |\widehat{\theta}(Y(j), Y(k)|Y(S))|. \tag{41}$$

Theorem 24. *Define Z_α by the bootstrap equation*

$$P\left(\max_{jk} \sqrt{n}|\widehat{\theta}^*_{jk} - \widehat{\theta}_{jk}| > Z_\alpha \ \Big| \ Y_1, \ldots, Y_n\right) = \alpha. \tag{42}$$

Let

$$\mathcal{R} = \left\{ a \in \mathbb{R}^{D^2} : \ ||a - \widehat{\theta}||_\infty \leq \frac{Z_\alpha}{\sqrt{n}} \right\}.$$

Then

$$P(\theta \in \mathcal{R}) \geq 1 - \alpha - O\left(\frac{(\log L)}{n^{1/8}}\right) - O\left(\left(\frac{\log L}{n}\right)^{1/6}\right).$$

The proof is basically the same as the proof of Theorem 21. We remark, however, that in this case, L has to be fixed and chosen in advance.

We think that the restricted partial correlation idea is very promising but currently we have no efficient way to compute the graph this way. To compute the restricted partial correlation we would need to do the following: for each pair (j, k) we have to search over the $\binom{D-2}{L}$ subsets and find the maximum. This is repeated for all D^2 pairs. Then the entire procedure needs to be bootstrapped. Despite the fact that the method is currently not computationally feasible, we include it because we believe that it may be possible in the future to find efficient computational approximations.

8. Experiments

In this section we illustrate the methods with some simple examples. We consider three models:

(B1) Dense Model: $\Omega_{jk} = a$ for all $j \neq k$.
(B2) Markov Chain: $Y(j) = aY(j+1) + \epsilon_j$.
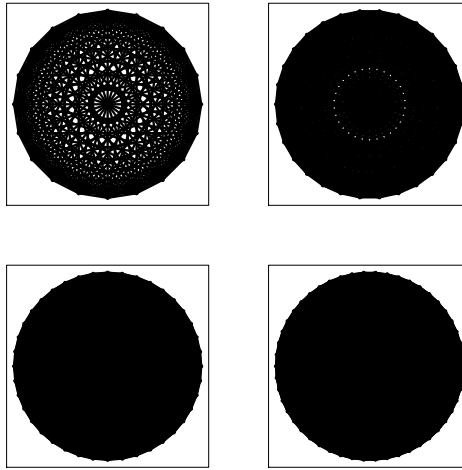(B3) Structural Equation Model: $Y(j) = a \sum_{s=1}^{j-1} Y(s) + \epsilon_j$, $j = 2, \ldots, D$.

FIG 3. *Bootstrap based undirected graph for Dense model with $\alpha = .9$, $a = .9$, $n = 100$ and dimensions 20,30,40,50.*
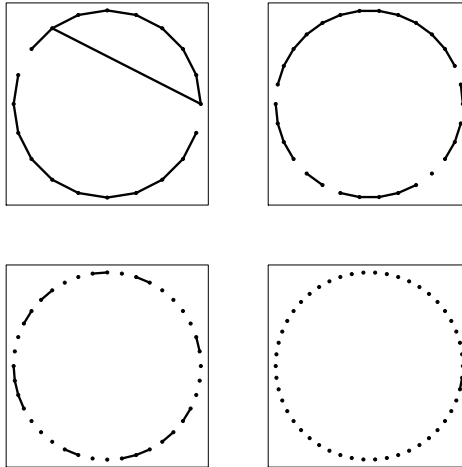

FIG 4. *Bootstrap based undirected graph for Markov model with $\alpha = .9$, $a = .9$, $n = 100$ and dimensions 20,30,40,50.*

The purpose of the experiments is to get some intuitive sense of how much information in the original graph is captured in the dimension reduced graph.

In each case we show results for the bootstrap. We stopped when the results became numerically unstable. Then we increased the dimension and switched to the high dimensional methods, namely, the cluster graphs, the correlation graphs and the restricted graphs. (We do not include the block graphs which did not work well.) The results are in Figures 3, 4, 5, 6, 7 and 8.
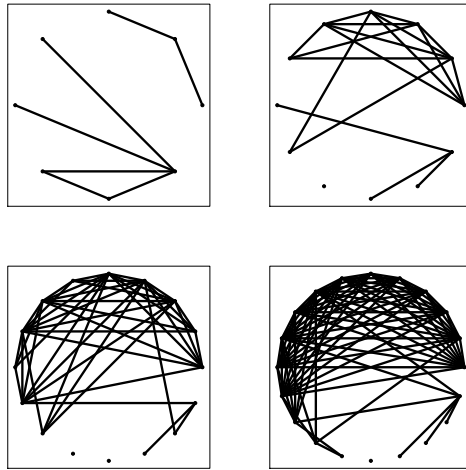
FIG 5. *Bootstrap based undirected graph for SEM model with $\alpha = .9$, $a = .5$, $n = 100$ and dimensions 8,12,16,20.*
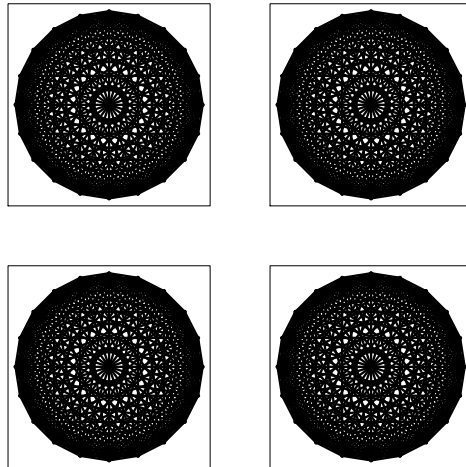
FIG 6. *Cluster graph for Dense model with $\alpha = .9$, $a = .9$, $n = 100$ and dimensions 70, 80, 90, 100 and $L = 20$.*

The results for the dense model are good up to $D = 50$. After that, the cluster graph method is used and it clearly captures the qualitative features of the graph. Validity holds as D increases but the power starts to decrease leading to missing edges. The cluster graph is interesting here as it obviously cannot reconstruct the Markov structure but still does capture interesting qualitative features of the underlying graph. The SEM model is difficult; it is a complete graph but some edges are harder to detect. The power again falls off as D increases. Again
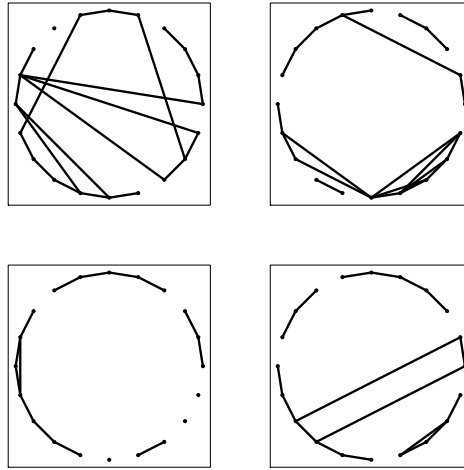
FIG 7. *Cluster graph for Markov model with* $\alpha = .9$, $a = .9$, $n = 100$ *and dimensions 70, 80, 90, 100 and* $L = 20$.
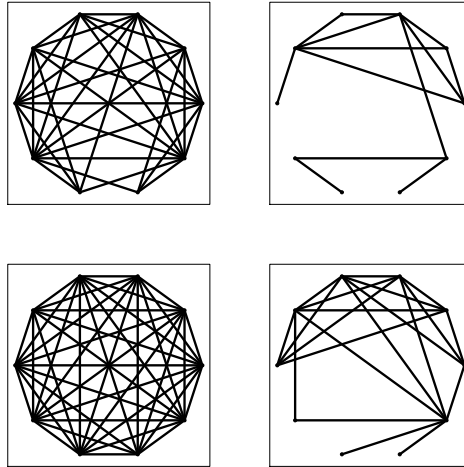

FIG 8. *Cluster graph for SEM model with* $\alpha = .9$, $a = .5$, $n = 100$ *and dimensions 28, 32, 36, 40 and* $L = 10$.

we see that the cluster graph loses information but permits us to find a graph with qualitative features similar to the true graph with higher dimensions.

The correlation graph for the dense and SEM models, while preserving validity has essentially no power. More precisely, the graphical model leaves a very small imprint in the correlation matrix. For example, the covariance in the dense model is easily seen to be $O(a/D)$. So while the inverse covariance matrix is dense, the covariance matrix has small entries. The correlation graph for the
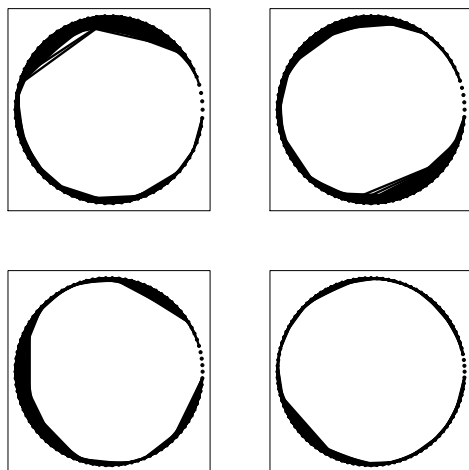
FIG 9. *Correlation graph for Markov model with $\alpha = .9$, $a = .9$, $n = 100$ and dimensions 70, 80, 90, 100 and $L = 20$.*
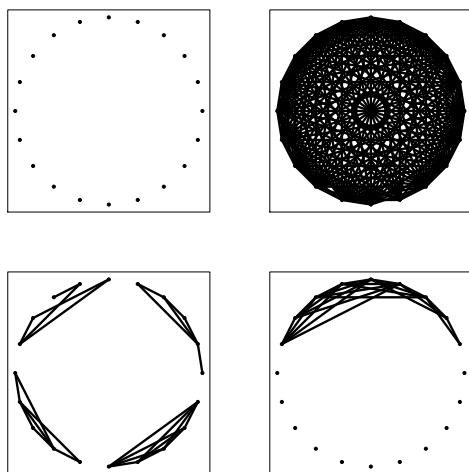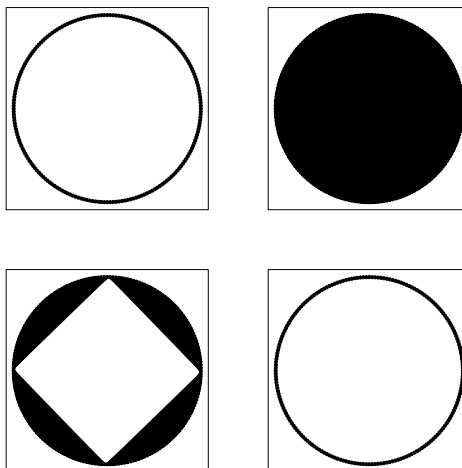

FIG 10. *Correlation Graphs, n=100, D=12.*

Markov model does contain useful information as shown in Figure 9. Of course, there are extra edges due to the induced correlations. Nevertheless, most of the essential structure is apparent.

We also considered the behavior of the correlation graph for a few other models. Figure 10 shows the correlation graph for a null model, a dense covariance matrix, a four-block model and a partial Markov chain (10 edges). In each case, $n = 100$ and $D = 12$. Figure 11 shows the same models but with $D = 200$. For these models the method does very well even with $D > n$.

FIG 11. *Correlation Graphs, n=100, D=200.*

As mentioned earlier, the restricted partial correlation graph is so computationally intensive that it is not yet practical. We believe the method is promising which is why we have included it in the paper but at this point we do not have numerical experiments.

Finally, as a sanity check, we checked the coverage of the bootstrap for two models: the null model (no edges) and the Markov model. We declare an error if there is even a single wrong edge. Using $\alpha = .10$ and $n = 100$ we have the following error rates:

Model/Dimension	D = 20	D = 50
Null	.01	.01
Markov	.00	.01

The error rates is well under α. Indeed, we see that the coverage is conservative as we would expect.

We re-ran all the experiments with $\alpha = 0.05$ and $\alpha = 0.01$. As expected, with smaller α we tend to get slightly sparser graphs. However, the differences were quite small. The cluster graph for the SEM model showed the greatest sensitivity to the choice of α. The others were quite insensitive.

9. Conclusion

We have described methods for inferring graphs that use weak assumptions and that have confidence guarantees. Our methods are atavistic: we use very traditional ideas that have been swept aside in light of the newer sparsity-based approaches. We do not mean in any way to criticize sparsity-based methods which we find fascinating. But our main message is that the older methods still have a role to play especially if we want methods that use weaker assumptions.

There are several open problems that we will address in the future. We briefly describe a few here. First, we do not have any theory to characterize how the original graph relates to the graph of the dimension reduced problem. It would be useful to have some general theory which shows which features of the original graph are preserved.

Perhaps the most important extension is to go beyond linear measures of dependence. Following Bergsma (2011), write

$$Y = g(X) + \epsilon_Y \quad \text{and} \quad Z = h(X) + \epsilon_Z$$

and define the nonparametric partial correlation

$$\theta_{YZ.X} = \frac{\mathbb{E}(\epsilon_Y \epsilon_Z)}{\sqrt{\mathbb{E}(\epsilon_Y^2)\mathbb{E}(\epsilon_Z^2)}}.$$

Let

$$\widehat{\epsilon}_{Y_i} = Y_i - \widehat{g}(X_i) \quad \text{and} \quad \widehat{\epsilon}_{Z_i} = Y_i - \widehat{h}(X_i).$$

Let

$$\widehat{\theta}_{YZ.X} = \frac{\sum_i \widehat{\epsilon}_{Y_i} \widehat{\epsilon}_{Z_i}}{\sqrt{\sum_i \widehat{\epsilon}_{Y_i}^2 \sum_i \widehat{\epsilon}_{Z_i}^2}}.$$

Bergsma shows that, for some $q_1, q_2 > 0$,

$$\sqrt{n}(\widehat{\theta}_{YZ.X} - \theta_{YZ.X}) = \sqrt{n}(r_{YZ.X} - \theta_{YZ.X}) + O_P\left(n^{-\min(q_1,q_2)}\right)$$

where

$$r_{YZ.X} = \frac{\sum_i \epsilon_{Y_i} \epsilon_{Z_i}}{\sqrt{\sum_i \epsilon_{Y_i}^2 \sum_i \epsilon_{Z_i}^2}}$$

and

$$n^{q_1}(\widehat{g}(x) - g(x)) = O_P(1), \quad n^{q_2}(\widehat{h}(x) - h(x)) = O_P(1).$$

One can then extend the techniques in this paper to get confidence measures.

Other problems for future development are: the development of computationally efficient methods for computing the restricted partial correlation graph and the extension of our theory to shrinkage graphs.

Appendix: Alternative delta method

If one is only interested in a single partial correlation, then one can use use a Taylor series together with the Berry-Esseen theorem. We provide this analysis here. At the end, we can turn this into a joint confidence set for all partial correlations using the union bound but this leads to a larger error than our earlier analysis. So the main interest of this section is single partial correlations.

Let us write $\theta_{jk} = g_{jk}(\sigma)$ where $g_{jk} : \mathbb{R}^{D \times D} \to [-1, 1]$. Let ℓ_{jk} and H_{jk} denote the gradient and Hessian of g_{jk}. Both ℓ_{jk} and H_{jk} are bounded continuous functions as long as Σ is invertible. The linearization of θ_{jk} is

$$\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk}) = \delta^T \ell_{jk} + \frac{R_{jk}}{\sqrt{n}} \tag{43}$$

where $\ell_{jk} \equiv \ell_{jk}(\sigma)$ and the remainder term R_{jk} is

$$R_{jk} = \frac{1}{2}\delta^T H_{jk}(\widetilde{\sigma})\delta \tag{44}$$

for some $\widetilde{\sigma}$ between σ and s. We compute ℓ_{jk} and H_{jk} explicitly in Section 6.5.

Let

$$s_{jk}^2 = U(\sigma), \quad \widehat{s}_{jk}^2 = U(s)$$

where

$$U_{jk}(\sigma) = \ell_{jk}(\sigma)^T T(\sigma)\ell_{jk}(\sigma). \tag{45}$$

The asymptotic variance of the linearized partial correlation $\delta^T \ell_{jk}$ is s_{jk}^2 and its estimate is $\widehat{s}_{jk}^2$.

Define $B = \{a : \ \|a - \sigma\| \leq C\sqrt{D^2 \log n/n}\}$. It follows from Lemma 7 that, for large enough C, $s \in B$ except on a set of probability at most $1/n$. Let

$$\xi_n = \sup_{a \in B} \max_{jk} \|\ell_{jk}(a)\|_1$$

$$\gamma_n = \sup_{a \in B} \max_{jk} \sqrt{\frac{\|H_{jk}(a)\|}{s_{jk}(a)}}$$

$$\rho_n = \sup_{a \in B} \max_{jk} \frac{\|Q'_{jk}(a)\|_1}{s_{jk}}.$$

Note that these constants are also functions of D.

We begin by approximating the distribution of a single partial correlation. Let

$$T_{jk} = \frac{\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk})}{s_{jk}}.$$

We start by assuming that $s_{jk}^2 = \ell_{jk}(\sigma)^T T(\sigma)\ell_{jk}(\sigma)$ is known.

Lemma 25. *We have*

$$\max_{j,k} \sup_z |P(T_{jk} \leq z) - \Phi(z)| \preceq \frac{1}{\sqrt{n}} + \frac{2\gamma_n}{\sqrt{n}} \log(nD^2).$$

Proof. We have

$$T_{jk} = \frac{U}{s_{jk}} + \frac{R_{jk}}{s_{jk}\sqrt{n}}$$

where $U = \sqrt{n}a^T(s-\sigma) = n^{-1}\sum_i V_i$ where $V_i = \text{vec}(Y_iY_i^T) - \sigma$ and $a = \ell_{jk}$. By Lemma 8, for every $\epsilon > 0$,

$$\sup_z |P(T_{jk} \le z) - \Phi(z)| \le \sup_z \left| P\left(\frac{U}{s_{jk}} \le z\right) - \Phi(z)\right| + \epsilon + P\left(\left|\frac{R_{jk}}{s_{jk}\sqrt{n}}\right| > \epsilon\right).$$

Note that $\text{Var}(V_i) = s_{jk}^2$ and

$$\mathbb{E}|V_i|^3 \le C\sum_i |a_j|^3.$$

Let $Z \sim N(0,1)$. By the Berry-Esseen theorem,

$$\sup_t \left| \mathbb{P}\left(\frac{U_n}{s_{jk}} \le t\right) - P(Z \le t)\right| \preceq \frac{\sum_j |a_j|^3}{\sqrt{n}(a^TTa)^{3/2}} \le \frac{\sum_j |a_j|^3}{\sqrt{n}c_0^{3/2}||a||^3} \le \frac{1}{\sqrt{n}}$$

since $||a||_3 \le ||a||_2$ and $\frac{\sum_j |a_j|^3}{||a||^3} = \frac{||a||_3^3}{||a||_2^3}$. Now

$$\left|\frac{R_{jk}}{s_{jk}\sqrt{n}}\right| = \frac{1}{2}\frac{\delta^T H_{jk}\delta}{s_{jk}\sqrt{n}} \le \frac{\gamma_n||\delta||_{\max}^2}{\sqrt{n}}.$$

From Lemma 8,

$$P\left(\left|\frac{R_{jk}}{s_{jk}\sqrt{n}}\right| > \epsilon\right) \le P\left(\frac{\gamma_n||\delta||_{\max}^2}{\sqrt{n}} > \epsilon\right) = P(||s-\sigma||_\infty > \frac{\sqrt{\epsilon}}{n^{1/4}\sqrt{\gamma}})$$
$$\le D^2 e^{-n\epsilon/(\gamma\sqrt{n})}.$$

Let $\epsilon = \frac{\gamma}{\sqrt{n}}\log(nD^2)$. Then $D^2 e^{-n\epsilon/(\gamma\sqrt{n})} \le \epsilon$. The result follows. $\square$

Now let
$$Z_{jk} = \frac{\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk})}{\widehat{s}_{jk}}$$

where $\widehat{s}_{jk}^2 = \ell_{jk}(s)^T T(s)\ell_{jk}(s)$.

Theorem 26.

$$\max_{j,k}\sup_z \left| P\left(\frac{\sqrt{n}(\widehat{\theta}_{jk}-\theta_{jk})}{\widehat{s}_{jk}} \le z\right) - \Phi(z)\right| \preceq \sqrt{\frac{\rho_n}{n}}\log(nD^2) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2).$$

Proof. Let $E = \{s_{jk}/\widehat{s}_{jk} > 1+\epsilon\}$ and $F = \{T_{jk} > u/\epsilon\}$ where $\epsilon = \sqrt{\rho_n/n}\log(nD^2)$ and $u = \epsilon\log(n)$. Note that $s_{jk} - \widehat{s}_{jk} = U(\sigma) - U(s) = (\sigma-s)^T Q'$ where Q' is the gradient of Q evaluated at some point between s and σ. Then, for $0 < \epsilon \le 1$,

$$P(E^c) = P\left(\frac{s_{jk} - \widehat{s}_{jk}}{s_{jk}} > \frac{\epsilon}{1+\epsilon}\right) = P\left(\frac{U(\sigma)-U(s)}{s_{jk}} > \frac{\epsilon}{1+\epsilon}\right)$$

$$= P\left(\frac{(\sigma - s)^T Q'}{s_{jk}} > \frac{\epsilon}{1+\epsilon}\right) \leq P\left(\frac{||s - \sigma||_\infty ||Q'||_1}{s_{jk}} > \frac{\epsilon}{1+\epsilon}\right)$$

$$\leq P\left(||s - \sigma||_\infty \rho_n > \frac{\epsilon}{1+\epsilon}\right) = P\left(||s - \sigma||_\infty > \frac{\epsilon}{2\rho_n}\right)$$

$$\leq D^2 e^{-n\epsilon^2/(4\rho_n^2)} \leq \epsilon.$$

Now,

$$P\left(\frac{\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk})}{\widehat{s}_{jk}} \leq z\right) - \Phi(z) = P\left(T_{jk}\left(\frac{s_{jk}}{\widehat{s}_{jk}}\right) \leq z\right) - \Phi(z)$$

$$\leq P\left(T_{jk}(1-\epsilon) \leq z\right) + P(E^c) - \Phi(z)$$
$$= P\left(T_{jk} - T_{jk}\epsilon \leq z\right) + P(E^c) - \Phi(z)$$
$$\leq P\left(T_{jk} \leq z + u\right) + P(F^c) + P(E^c) - \Phi(z)$$
$$\leq P\left(T_{jk} \leq z + u\right) - \Phi(z+u) + P(F^c) + P(E^c) + u$$
$$\leq P\left(T_{jk} \leq z + u\right) - \Phi(z+u) + P(F^c) + \epsilon + u.$$

Now

$$P(F^c) = P(T_{jk} > u/\epsilon) \leq P(Z > u/\epsilon) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2)$$

$$= P(Z > \log n) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2)$$

$$\preceq \frac{\gamma_n}{\sqrt{n}}\log(nD^2).$$

So,

$$P\left(\frac{\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk})}{\widehat{s}_{jk}} \leq z\right) - \Phi(z) \leq P\left(T_{jk} \leq z + u\right) - \Phi(z+u) + \epsilon + u + \frac{1}{n}$$

$$+ \frac{\gamma_n}{\sqrt{n}}\log(nD^2)$$

$$\preceq \sqrt{\frac{\rho_n}{n}}\log(nD^2) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2).$$

Taking the supremum over z gives an upper. A similar lower bound completes the proof. □

Now we turn to bounding $P(\max_{jk}|Z_{jk}| > z)$. We use the union bound. So,

$$P(\max_{jk}|Z_{jk}| > z) \leq \sum_{jk} P(|Z_{jk}| > z)$$

$$= D^2\Phi(z) + \sum_{jk}[P(|Z_{jk}| > z) - \Phi(z)]$$

$$\leq D^2\Phi(z) + D^2\left[\sqrt{\frac{\rho_n}{n}}\log(nD^2) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2)\right].$$

Setting $z = -\Phi(\alpha/D^2)$ we have that

$$P(\max_{jk}|Z_{jk}| > z) \le \alpha + D^2\left[\sqrt{\frac{\rho_n}{n}}\log(nD^2) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2)\right].$$

Corollary 27. *Let $z \equiv z_{\alpha/D^2}$ and let*

$$R = \bigotimes_{j,k}\left[\widehat{\theta}_{jk} - \frac{z\widehat{s}_{jk}}{\sqrt{n}},\ \widehat{\theta}_{jk} + \frac{z\widehat{s}_{jk}}{\sqrt{n}}\right].$$

Then

$$P(\theta \in R) = 1 - \alpha + D^2\left[\sqrt{\frac{\rho_n}{n}}\log(nD^2) + \frac{\gamma_n}{\sqrt{n}}\log(nD^2)\right].$$

Note the presence of the D^2 term. This term is avoided in the analysis in Section 6.

References

BERGSMA, W. (2011). A note on the distribution of the partial correlation coefficient with nonparametrically estimated marginal regressions. *arXiv:1101.4616*.

BOIK, R. and HAALAND, B. (2006). Second-order accurate inference on simple, partial, and multiple correlations. *Journal of Modern Applied Statistical Methods* **5** 283–308.

CASTELO, R. and ROVERATO, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *The Journal of Machine Learning Research* **7** 2621–2650. MR2274453

CHEN, L. H. and SHAO, Q.-M. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli* 581–599. MR2331265

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2012). Central limit theorem and multiplier boostrap when p is much larger than n. *arXiv:1212.6906*. MR3156748

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *arXiv:1301.4807*. MR3161448

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. Springer Verlag.

DRTON, M. and PERLMAN, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91** 591–602. MR2090624

FRIEDMAN, J. and TIBSHIRANI, R. (2007). Graphical lasso.

HARRIS, N. and DRTON, M. (2012). PC algorithm for Gaussian copula graphical models. *arXiv:1207.0242*.

HORN, R. A. and JOHNSON, C. R. (1990). *Matrix analysis*. Cambridge university press. MR1084815

JALALI, A., JOHNSON, C. C. and RAVIKUMAR, P. D. (2011). On Learning Discrete Graphical Models using Greedy Methods. In *NIPS* 1935–1943.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88** 365–411. MR2026339

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *arXiv:1306.0976*. MR3161453

MAGNUS, X. and NEUDECKER, H. (1988). *Matrix differential calculus*. New York.

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* 255–285. MR1212176

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462. MR2278363

PINELIS, I. and MOLZON, R. (2013). Berry-Esseen bounds for general nonlinear statistics, with applications to Pearson's and non-central Student's and Hotelling's. *arXiv:0906.0177*.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics* **16** 356–366. MR0924876

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G., YU, B. et al. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980. MR2836766

REN, Z., SUN, T., ZHANGE, C.-H. and ZHOU, H. (2013). Asymptotic normality and optimalities in estimation of large Gaissian graphical models. *Manuscript*.

SCHÄFER, J., STRIMMER, K. et al. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4** 32.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*. MR2963170

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824