The Annals of Statistics
2018, Vol. 46, No. 6B, 3422–3450
https://doi.org/10.1214/17-AOS1663

ROCKET: ROBUST CONFIDENCE INTERVALS VIA KENDALL'S TAU FOR TRANSELLIPTICAL GRAPHICAL MODELS[1]

BY RINA FOYGEL BARBER AND MLADEN KOLAR

University of Chicago

Understanding complex relationships between random variables is of fundamental importance in high-dimensional statistics, with numerous applications in biological and social sciences. Undirected graphical models are often used to represent dependencies between random variables, where an edge between two random variables is drawn if they are conditionally dependent given all the other measured variables. A large body of literature exists on methods that estimate the structure of an undirected graphical model, however, little is known about the distributional properties of the estimators beyond the Gaussian setting. In this paper, we focus on inference for edge parameters in a high-dimensional transelliptical model, which generalizes Gaussian and nonparanormal graphical models. We propose ROCKET, a novel procedure for estimating parameters in the latent inverse covariance matrix. We establish asymptotic normality of ROCKET in an ultra high-dimensional setting under mild assumptions, without relying on oracle model selection results. ROCKET requires the same number of samples that are known to be necessary for obtaining a $\sqrt{n}$ consistent estimator of an element in the precision matrix under a Gaussian model. Hence, it is an optimal estimator under a much larger family of distributions. The result hinges on a tight control of the sparse spectral norm of the nonparametric Kendall's tau estimator of the correlation matrix, which is of independent interest. Empirically, ROCKET outperforms the nonparanormal and Gaussian models in terms of achieving accurate inference on simulated data. We also compare the three methods on real data (daily stock returns), and find that the ROCKET estimator is the only method whose behavior across subsamples agrees with the distribution predicted by the theory.

1. Introduction. Probabilistic graphical models [Lauritzen (1996)] have been widely used to explore complex systems and aid scientific discovery in areas ranging from biology and neuroscience to financial modeling and social media analysis. An undirected graphical model consists of a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of vertices and E is the set of edges, and a p-dimensional random vector $X = (X_1, \ldots, X_p)^\top$ that is Markov with respect to G. In particular, we have that X_a and X_b are conditionally independent given the remaining

Received February 2016; revised April 2017.

[1]Supported in part by an IBM Corporation Faculty Research Fund at the University of Chicago Booth School of Business, and an Alfred P. Sloan Fellowship.

MSC2010 subject classifications. Primary 62G10; secondary 62F12, 62G20.

Key words and phrases. Graphical model selection, transelliptical graphical models, covariance selection, uniformly valid inference, post-model selection inference, rank-based estimation.

variables $\{X_c \mid c \in \{1, \ldots, p\} \backslash \{a, b\}\}$ if and only if $\{a, b\} \notin E$. One of the central questions in high-dimensional statistics is estimation of the undirected graph G given n independent realizations of X, as well as quantifying uncertainty of the estimator.

In this paper, we focus on (asymptotic) inference for elements in the latent inverse covariance matrix under the semiparametric elliptical copula model [Embrechts, Lindskog and McNeil (2003), Klüppelberg, Kuhn and Peng (2008)], also known as the transelliptical model [Liu, Han and Zhang (2012)]. Let $X_1, \ldots, X_n$ be n independent copies of the random vector X that follows a transelliptical distribution,

$$X \sim \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_p), \tag{1.1}$$

where $\Sigma \in \mathbb{R}^p$ is a correlation matrix (i.e., $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$), $\xi \in \mathbb{R}$ is a nonnegative random variable with $\mathbb{P}\{\xi = 0\} = 0$, and $f_1, \ldots, f_p$ are univariate, strictly increasing functions. Recall that X follows a transelliptical distribution if the marginal transformation $(f_1(X_1), \ldots, f_p(X_p))$ of X follows a (centered) elliptically contoured distribution with covariance matrix Σ [Fang, Kotz and Ng (1990)]. Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, also known as the precision matrix. Under a Gaussian model, nonzero elements in Ω correspond to pairs of variables that are conditionally dependent, that is, form an edge in the graph G; under an elliptical model, nonzero elements in Ω correspond to variables that are conditionally correlated [but in general it is possible to have $\Omega_{ab} = 0$ where $f_a(X_a)$ and $f_b(X_b)$ are conditionally uncorrelated, but not conditionally independent]. Under the model in (1.1), we construct an estimator for a fixed element of the precision matrix, Ω_{ab}, that is asymptotically normal. Furthermore, we construct a confidence interval for the unknown parameter Ω_{ab} that is valid and robust to model selection mistakes. Finally, we construct a uniformly valid hypothesis test for the presence of an edge in the graphical model.

Our main theoretical result establishes that given initial estimates of the regression coefficients for $(f_a(X_a), f_b(X_b)$ on $(f_j(X_j))_{j \neq a,b}$, one can obtain a $\sqrt{n}$-consistent and asymptotically normal estimator for Ω_{ab}. These initial estimators need to converge at a sufficiently fast rate (see Section 3). In particular, we note that we do not require strict sparsity in these regressions, and allow for an error rate that is achievable by known methods such as a nonconvex Lasso [Loh and Wainwright (2015)] (see Section 3.1). To achieve $\sqrt{n}$-consistent rate, our estimator requires the same scaling for the sample size n as in the Gaussian case; this sample size scaling is minimax optimal [Ren et al. (2015)].

Given accurate initial estimates, in order to construct the asymptotically normal estimator, we prove a key result: that the vector $\operatorname{sign}(X_i - X_{i'})$ is sub-Gaussian at the scale $\mathsf{C}(\Sigma)$ (the condition number of Σ), with dependence on the dimension p coming only through $\mathsf{C}(\Sigma)$ (this problem was initially posed by Han and Liu (2013), where sub-Gaussianity was proved for some special cases). This result

allows us to construct an asymptotically normal estimator by combining the initial regression coefficient estimates with the Kendall's tau rank correlation matrix. In particular, the sub-Gaussianity result allows us to establish a new concentration result on the operator norm of the Kendall's tau correlation matrix that holds with exponentially high probability. This result allows us to uniformly control deviations of quadratic forms involving the Kendall's tau correlation matrix over approximately sparse vectors. These results are of independent interest and could be used to extend recent results of Mitra and Zhang (2014), Wegkamp and Zhao (2016) and Han and Liu (2013) to the elliptical copula setting. Furthermore, sub-Gaussianity of $\operatorname{sign}(X_i - X_{i'})$, which in turn leads to a bound on the error of the Kendall's tau estimate of Σ in the sparse spectral norm, allows us to study properties of penalized rank regression in high dimensions.

We base our confidence intervals and hypothesis tests on the asymptotically normal estimator of the element Ω_{ab} (see Section 2). We point out that our results hold under milder conditions than those required in Ren et al. (2015), which treats the special case of Gaussian graphical models. Most notably, we give a $\sqrt{n}$-consistent estimator for elements in the precision matrix without requiring strong parametric assumptions.

1.1. *Relationship to literature.* Our work contributes to several areas. First, we contribute to the growing literature on graphical model selection in high dimensions. There is extensive literature on the Gaussian graphical model, where it is assumed that $X \sim N(0, \Sigma)$, in which case the edge set E of the graph G is encoded by the nonzero elements of the precision matrix Ω [Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Rothman et al. (2008), Friedman, Hastie and Tibshirani (2008), d'Aspremont, Banerjee and El Ghaoui (2008), Fan, Feng and Wu (2009), Lam and Fan (2009), Yuan (2010), Cai, Liu and Luo (2011), Liu and Wang (2017), Zhao and Liu (2014)]. Learning structure of the Ising model based on the penalized pseudo-likelihood was studied in Höfling and Tibshirani (2009), Ravikumar, Wainwright and Lafferty (2010) and Xue, Zou and Cai (2012). More recently, Yang et al. (2015) studied estimation of graphical models under the assumption that each of the nodes' conditional distribution belongs to an exponential family distribution. See also Guo et al. (2011a, 2011b), Lee and Hastie (2012), Cheng et al. (2017), Yang et al. (2012) and Yang et al. (2014) who studied mixed graphical models, where the nodes' conditional distributions are not necessarily all from the same family (for instance, there may be continuous-valued nodes as well as discrete-valued nodes). The parametric Gaussian assumption was relaxed in Liu, Lafferty and Wasserman (2009), where graph estimation was studied under a Gaussian copula model. More recently, Liu et al. (2012), Xue and Zou (2012) and Liu, Han and Zhang (2012) show that the graph can be recovered in the Gaussian and elliptical semiparametric model class under the same conditions on the sample size n, number of nodes p and the maximum node degree in the graph k as if the estimation was done under the Gaussian assumption. In our paper, we

construct a novel $\sqrt{n}$-consistent estimator of an element in the precision matrix without requiring oracle model selection properties.

Second, we contribute to the literature on high-dimensional inference. Recently, there has been much interest on performing valid statistical inference in the high-dimensional setting. Zhang and Zhang (2014), Belloni, Chernozhukov and Hansen (2014), Belloni, Chernozhukov and Wei (2013), van de Geer et al. (2014), Javanmard and Montanari (2014), Javanmard and Montanari (2013) and Farrell (2015) developed methods for construction of confidence intervals for low-dimensional parameters in high-dimensional linear and generalized linear models, as well as hypothesis tests. These methods construct honest, uniformly valid confidence intervals and hypothesis tests based on the ℓ_1-penalized estimator in the first stage. Similar results were obtained in the context of the ℓ_1-penalized least absolute deviation and quantile regression [Belloni, Chernozhukov and Kato (2013a, 2013b)]. Lockhart et al. (2014) study significance of the input variables that enter the model along the lasso path. Lee et al. (2016) and Tibshirani et al. (2016) perform post-selection inference conditional on the selected model. Liu (2013), Ren et al. (2015) and Chen et al. (2016) construct $\sqrt{n}$-consistent estimators for elements of the precision matrix Ω under a Gaussian assumption. We extend these results to perform valid inference under semiparametric elliptical copula models. In a recent independent work, Gu et al. (2015) propose a procedure for inference under a nonparanormal model. We will provide a detailed comparison in Section 3 and Section 5.

1.2. *Notation.* Let $[n]$ denote the set $\{1, \ldots, n\}$ and let $\mathbb{1}\{\cdot\}$ denote the indicator function. For a vector $a \in \mathbb{R}^d$, we let $\operatorname{supp}(a) = \{j : a_j \neq 0\}$ be the support set, and let $\|a\|_q$, for $q \in [1, \infty)$, be the ℓ_q-norm defined as $\|a\|_q = (\sum_{i\in[n]} |a_i|^q)^{1/q}$ with the usual extensions for $q \in \{0, \infty\}$, that is, $\|a\|_0 = |\operatorname{supp}(a)|$ and $\|a\|_\infty = \max_{i\in[n]} |a_i|$.

For a matrix $A \in \mathbb{R}^{n_1\times n_2}$, for sets $S \subset [n_1]$ and $T \subset [n_2]$, we write A_{ST} to denote the $|S| \times |T|$ submatrix of A obtained by extracting the appropriate rows and columns. The sets S and/or T can be replaced by single indices, for example, for $S \subset [n_1]$ and $j \in [n_2]$, A_{Sj} is a $|S|$-length vector. If $A \in \mathbb{R}^{n\times n}$ is a square matrix, for any $T \subset [n]$ we may write A_T to denote the square submatrix A_{TT}.

For a matrix $A \in \mathbb{R}^{n_1\times n_2}$, we use the notation $\mathsf{vec}(A)$ to denote the vector in $\mathbb{R}^{n_1 n_2}$ formed by stacking the columns of A. We denote the Frobenius norm of A by $\|A\|_{\mathsf{F}}^2 = \sum_{i\in[n_1], j\in[n_2]} A_{ij}^2$, and the operator norm (spectral norm) by $\|A\|_{\mathsf{op}}$, that is, the largest singular value of A. The norms $\|A\|_1$ and $\|A\|_\infty$ are applied entrywise, with $\|A\|_1 = \sum_{ij} |A_{ij}|$ and $\|A\|_\infty = \max_{ij} |A_{ij}|$. We write $\mathsf{C}(A)$ to denote the condition number of A, that is, the ratio between the largest and smallest singular values. For two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{r\times s}$, $A \otimes B \in \mathbb{R}^{nr\times ms}$ denotes the Kronecker product, with $(A \otimes B)_{ik,jl} = A_{ij} B_{kl}$. For two matrices of the same size, $A, B \in \mathbb{R}^{n\times m}$, $A \circ B \in \mathbb{R}^{n\times m}$ denotes the Hadamard product (i.e., the

entrywise product), with $(A \circ B)_{ij} = A_{ij} B_{ij}$. Kronecker products and Hadamard products are defined also for vectors, by treating a vector as a matrix with one column.

Throughout, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, that is, $\Phi(t) = \mathbb{P}\{N(0,1) \leqslant t\}$.

2. Preliminaries and method. Before introducing our method, we begin with some preliminary definitions and properties of the transelliptical distribution and related models.

Gaussian and nonparanormal graphical models. Suppose that $X = (X_1, \ldots, X_p)$ follows a multivariate normal distribution, $X \sim N(\mu, \Sigma)$. A Gaussian graphical model represents the structure of the covariance matrix Σ with a graph, where an edge between nodes a and b indicates that $\Omega_{ab} \neq 0$, where $\Omega = \Sigma^{-1}$ is the precision (inverse covariance) matrix. This model can be generalized by allowing for arbitrary marginal transformations on the variables $X_1, \ldots, X_p$. Liu, Lafferty and Wasserman (2009) study the resulting distribution, the nonparanormal model (also known as a Gaussian copula), where we write $X \sim \mathsf{NPN}(\Sigma; f_1, \ldots, f_p)$, if the marginally transformed vector $(f_1(X_1), \ldots, f_p(X_p))$ follows a (centered) multivariate normal distribution,

$$\big(f_1(X_1), \ldots, f_p(X_p)\big) \sim N(0, \Sigma).$$

The sparse structure of the underlying graphical model, representing the sparsity pattern in $\Omega = \Sigma^{-1}$, can then be recovered using similar methods as in the Gaussian case. Note that the Gaussian model is a special case of the nonparanormal model (by setting $f_1, \ldots, f_p$ each to be the identity function, or to be linear functions if we would like a nonzero mean).

Elliptical and transelliptical graphical models. The elliptical model is a generalization of the Gaussian graphical model that allows for heavier-tailed dependence between variables. The random vector $X = (X_1, \ldots, X_p)$ follows an elliptical distribution with the mean vector $\mu \in \mathbb{R}^p$, covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ and a random variable (the "radius") $\xi \geqslant 0$, denoted by $X \sim \mathsf{E}(\mu, \Sigma, \xi)$, if we can write $X = \mu + \xi \cdot A \cdot U$, where $AA^\top = \Sigma$ is a Cholesky decomposition of Σ, and where $U \in \mathbb{R}^p$ is a unit vector drawn uniformly at random (independently from the radius ξ). Note that the level sets of this distribution are given by ellipses, centered at μ and with shape determined by Σ. The Gaussian model is a special case of the elliptical model (by taking $\xi \sim \chi_p$).

The transelliptical model (also known as an elliptical copula) combines the elliptical distribution with marginal transformations, much as the nonparanormal distribution applies marginal transformations to a multivariate Gaussian. For a random vector $X \in \mathbb{R}^p$, we write

$$X \sim \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_p)$$

to denote that the marginally transformed vector $(f_1(X_1), \ldots, f_p(X_p))$ follows a centered elliptical distribution, specifically,

$$\big(f_1(X_1), \ldots, f_p(X_p)\big) \sim \mathsf{E}(0, \Sigma, \xi).$$

Here, the marginal transformation functions $f_1, \ldots, f_p$ are assumed to be strictly increasing. Note that the Gaussian, nonparanormal and elliptical models are each special cases of this model.

Pearson's rho and Kendall's tau. From this point on, we assume for each distribution that $\mu = 0$ and that Σ is a correlation matrix (i.e., diagonal elements are equal to one, $\Sigma_{aa} = 1$). In the case of the Gaussian distribution $X \sim N(0, \Sigma)$, the entries of Σ are the (population-level) Pearson's correlation coefficients for each pair of variables, which in this case we can also write as $\Sigma_{ab} = \mathbb{E}[X_a X_b]$. In this setting, we can estimate Σ with the sample covariance.

In the nonparanormal setting, $X \sim \mathsf{NPN}(\Sigma; f_1, \ldots, f_p)$, it is no longer the case that Σ_{ab} is equal to the (population-level) correlation $\mathrm{Corr}(X_a, X_b)$, due to the marginal transformations. However, we can estimate $f_1, \ldots, f_p$ by performing marginal empirical transformations of each X_a to the standard normal distribution. After taking these empirical transformations, Σ can again be estimated via the empirical covariances. Similarly, for the elliptical model $X \sim \mathsf{E}(0, \Sigma, \xi)$, after rescaling so that $\mathbb{E}[\xi^2] = p$ we also have $\Sigma_{ab} = \mathbb{E}[X_a X_b]$. We can therefore again estimate Σ via the empirical covariance.

For the transelliptical distribution, in contrast, this is no longer possible. Taking scaling $\mathbb{E}[\xi^2] = p$ for simplicity, we generalize the calculations above to have $\Sigma_{ab} = \mathbb{E}[f_a(X_a) f_b(X_b)]$. Therefore, if we can estimate the marginal transformations $f_1, \ldots, f_p$, then we can estimate Σ using the empirical covariance of the transformed data. However, unlike the nonparanormal model, estimating $f_1, \ldots, f_p$ is not straightforward. The reason is that, for the elliptical distribution $\mathsf{E}(0, \Sigma, \xi)$, the marginal distributions are not known unless the distribution of the radius ξ is known. Therefore, marginally for each X_a, we cannot estimate f_a because we do not know what should be the marginal distribution after transformation, that is, what should be the marginal distribution of $f_a(X_a)$. [In contrast, in the nonparanormal model, $f_a(X_a)$ is marginally normal.]

As an alternative, Liu, Han and Zhang (2012) use the Kendall rank correlation coefficient (Kendall's tau). At the population level, it is given by

$$\tau_{ab} := \tau(X_a, X_b) = \mathbb{E}\big[\mathrm{sign}\big(X_a - X_a'\big) \cdot \mathrm{sign}\big(X_b - X_b'\big)\big],$$

where X' is an i.i.d. copy of X. Unlike Pearson's rho, the Kendall's tau coefficient is invariant to marginal transformations: since f_a, f_b are strictly increasing functions, we see that

$$\begin{aligned}&\mathrm{sign}\big(f_a(X_a) - f_a\big(X_a'\big)\big) \cdot \mathrm{sign}\big(f_b(X_b) - f_b\big(X_b'\big)\big)\\ &\quad = \mathrm{sign}\big(X_a - X_a'\big) \cdot \mathrm{sign}\big(X_b - X_b'\big).\end{aligned}$$

At the sample level, Kendall's tau can be estimated by taking a U-statistic comparing each pair of distinct observations:

$$\hat{\tau}_{ab} = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < i' \leqslant n} \operatorname{sign}(X_{ia} - X_{i'a}) \cdot \operatorname{sign}(X_{ib} - X_{i'b}). \tag{2.1}$$

When X follows an elliptical distribution, Theorem 2 of Lindskog, McNeil and Schmock (2003) gives us the following relationship between Kendall's tau and the Pearson's rho coefficients given by the covariance matrix Σ:

$$\Sigma_{ab} = \sin\left(\frac{\pi}{2}\tau_{ab}\right) \qquad \text{for each } a, b \in [p]. \tag{2.2}$$

Since Kendall's tau is invariant to marginal transformations, this identity holds for the transelliptical family as well. For this reason, Liu, Han and Zhang (2012) estimate the covariance matrix Σ by

$$\hat{\Sigma}_{ab} = \sin\left(\frac{\pi}{2}\hat{\tau}_{ab}\right). \tag{2.3}$$

Note, however, that $\hat{\Sigma}$ is not necessarily positive semidefinite.

While Spearman's rho, like Kendall's tau, is also invariant to marginal transformations, Liu, Han and Zhang (2012) comment that there is no equivalence between Σ and the population-level Spearman's rho values [analogous to (2.2) for Kendall's tau] which holds uniformly across the entire elliptical (or transelliptical) family. Therefore, this type of estimator as in (2.3) could only be carried out with Kendall's tau.

For the remainder of this paper, $\hat{\Sigma}$ denotes the estimate given here in (2.3). The matrix of the Kendall's tau coefficients is denoted as T, with entries $T_{ab} := \tau_{ab}$, and $\hat{T}$ denotes its empirical estimate [with entries as in (2.1)].

Comparing models: *Tail dependence*. It is clear that, compared to a Gaussian graphical model, the nonparanormal model allows for data that may be extremely heavy-tailed (in the marginal distributions). A more subtle consideration is the question of tail dependence between two or more of the variables. In particular, the nonparanormal model does not allow for tail dependence between two variables to be any stronger than in the Gaussian distribution itself. Specifically, consider pairwise α-tail dependence between X_a and X_b, given by

$$\mathsf{Tail}_\alpha(X_a, X_b) := \operatorname{Corr}\big(\mathbb{1}\{X_a \geqslant q_\alpha^{X_a}\}, \mathbb{1}\{X_b \geqslant q_\alpha^{X_b}\}\big),$$

where $q_\alpha^{X_a}$ is the α-quantile of the marginal distribution of X_a, and same for X_b. Taking $\alpha \to 1$, this is a measure of the correlation between the extreme right tail of X_a and the extreme right tail of X_b. (Of course, we can also consider the left tail of the distribution of X_a and/or X_b.)

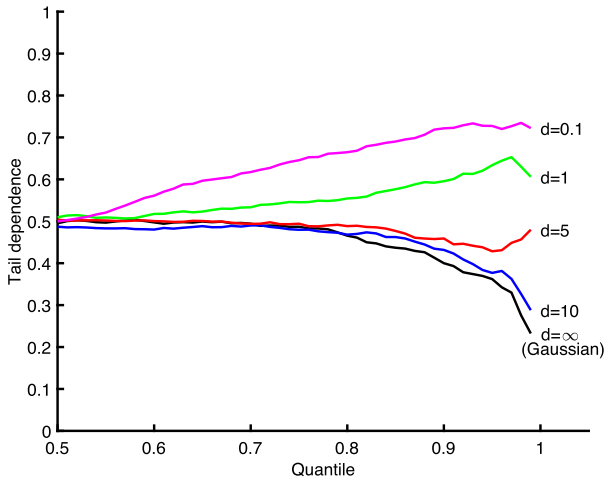

FIG. 1. *Tail dependence for normal and elliptical distributions on* $\mathbb{R}^2$. *Data is generated as in* (2.4). *The figure displays* $\mathsf{Tail}_\alpha(X_1, X_2)$, *estimated empirically from a sample size* $n = 20{,}000$.

Note that marginal transformations of each variable do not affect this measure, since the quantiles $q_\alpha^{X_a}, q_\alpha^{X_b}$ take these transformations into account. In particular, the nonparanormal distribution has the same tail correlations $\mathsf{Tail}_\alpha(X_a, X_b)$ as the multivariate Gaussian distribution (with the same Σ). In contrast, an elliptical or transelliptical model can exhibit much higher tail correlations. Since real data often exhibits heavy tail dependence between variables, the flexible transelliptical model may be a better fit in many applications.

We demonstrate this behavior with a simple example in Figure 1. Take

$$X = (X_1, X_2) \sim \mathsf{E}(0, \Sigma, \xi) \qquad \text{with } \Sigma = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{pmatrix}, \tag{2.4}$$

where $\xi \sim \chi_2 \cdot \sqrt{d}/\chi_d$ for $d \in \{0.1, 1, 5, 10, \infty\}$, corresponding to a multivariate t-distribution with d degrees of freedom [note that $d = \infty$ is equivalent to taking $X \sim N(0, \Sigma)$]. Note that at $\alpha = 0.5$, the relevant quantiles are $q_\alpha^{X_1} = q_\alpha^{X_2} = 0$, and so the tail correlation $\mathsf{Tail}_\alpha(X_1, X_2)$ is equal to the Kendall's tau coefficient $\tau(X_1, X_2) = \frac{2}{\pi}\arcsin(\Sigma_{12}) = 0.5$ at any value of d. Figure 1 shows that, as $\alpha \to 1$, the tail correlation decreases toward zero for the normal distribution ($d = \infty$) but grows for low values of d.

Therefore, the shift from a nonparanormal to a transelliptical model is important, since it allows us to model variables with high tail dependence, that is, high dependence between their "extreme events."

2.1. *ROCKET: An asymptotically normal estimator.* Suppose that our data points X_i are drawn i.i.d. from a transelliptical distribution with covariance matrix Σ. We would like to perform inference on a particular entry of the precision

matrix $\Omega = \Sigma^{-1}$, specifically, we are interested in producing a confidence interval for Ω_{ab} where $a \neq b \in \{1, \ldots, p\}$ is a prespecified node pair.

To move toward constructing a confidence interval, we introduce a few definitions and calculations. First, let $I = \{1, \ldots, p\} \backslash \{a, b\}$, and observe that by blockwise matrix inversion, we can calculate the $\{a, b\} \times \{a, b\}$ sub-block of Ω as follows:

$$\Omega_{ab,ab} = \left(\Sigma_{ab,ab} - \Sigma_{ab,I}\Sigma_I^{-1}\Sigma_{I,ab}\right)^{-1}. \tag{2.5}$$

Define $\gamma_a = \Sigma_I^{-1}\Sigma_{Ia}$ and $\gamma_b = \Sigma_I^{-1}\Sigma_{Ib}$. In the nonparanormal graphical model setting, these are the regression coefficients when $f_a(X_a)$ or $f_b(X_b)$ is regressed on $\{f_j(X_j) : j \in I\}$; in the linear model setting, this idea has been used in Sun and Zhang (2012a) and Belloni, Chernozhukov and Hansen (2014). We then have

$$\Sigma_{ab,I}\Sigma_I^{-1}\Sigma_{I,ab} = (\gamma_a\gamma_b)^\top \Sigma_{I,ab} = \Sigma_{I,ab}^\top(\gamma_a\gamma_b) = (\gamma_a\gamma_b)^\top \Sigma_I(\gamma_a\gamma_b).$$

We can therefore rewrite (2.5) as follows (this somewhat redundant formulation will allow for a favorable cancellation of error terms later on):

$$\begin{aligned}\Theta &:= (\Omega_{ab,ab})^{-1} \\ &= \Sigma_{ab,ab} - (\gamma_a\gamma_b)^\top \Sigma_{I,ab} - \Sigma_{I,ab}^\top(\gamma_a\gamma_b) + (\gamma_a\gamma_b)^\top \Sigma_I(\gamma_a\gamma_b).\end{aligned} \tag{2.6}$$

We abuse notation and index the entries of Θ with the indices a and b, that is, we denote Θ as lying in $\mathbb{R}^{\{a,b\}\times\{a,b\}}$ rather than $\mathbb{R}^{2\times 2}$.

Next, we define an oracle estimator of Θ, defined by plugging the *true* values of γ_a and γ_b and the *empirical* estimate of Σ [given in (2.3)] into (2.6) above:

$$\widetilde{\Theta} = \widehat{\Sigma}_{ab,ab} - (\gamma_a\gamma_b)^\top \widehat{\Sigma}_{I,ab} - \widehat{\Sigma}_{I,ab}^\top(\gamma_a\gamma_b) + (\gamma_a\gamma_b)^\top \widehat{\Sigma}_I(\gamma_a\gamma_b). \tag{2.7}$$

Later on (in Theorem 4.1), we will show that due to standard results on the theory of U-statistics, this oracle estimator is asymptotically normal. If $\widetilde{\Theta}$ were known, then we would have achieved our goal for inference in this model, as $\tilde{\Omega}_{ab} = (\tilde{\Theta}^{-1})_{ab}$ weakly converges to a normal random variable centered at Ω_{ab} with variance that scales as $\mathcal{O}(1/n)$ (we calculate this variance later).

Of course, in practice we do not know the true values of γ_a and γ_b, and must instead use some available estimators, denoted by $\breve{\gamma}_a$ and $\breve{\gamma}_b$ (we discuss how to obtain these preliminary estimates later on). Given the estimators of the regression vectors, we then define our estimator of Θ as follows:

$$\breve{\Theta} = \widehat{\Sigma}_{ab,ab} - (\breve{\gamma}_a\breve{\gamma}_b)^\top \widehat{\Sigma}_{I,ab} - \widehat{\Sigma}_{I,ab}^\top(\breve{\gamma}_a\breve{\gamma}_b) + (\breve{\gamma}_a\breve{\gamma}_b)^\top \widehat{\Sigma}_I(\breve{\gamma}_a\breve{\gamma}_b). \tag{2.8}$$

Since we are interested in Ω_{ab} rather than in the matrix Θ, as a final step we define our estimator

$$\breve{\Omega}_{ab} = \left(\breve{\Theta}^{-1}\right)_{ab}. \tag{2.9}$$

In order to make inference about Ω_{ab}, we approximate the distribution of $\breve{\Omega}_{ab}$, which is a function of $\breve{\Theta}$. We first treat the distribution of the corresponding entry in the oracle estimator $\widetilde{\Theta}$. To do so, let $u, v \in \mathbb{R}^{p_n}$ be the vectors with entries

$$u_a = 1, \qquad u_b = 0, \qquad u_I = -\gamma_a \quad \text{and} \quad v_a = 0, \qquad v_b = 1, \qquad v_I = -\gamma_b,$$

and observe from (2.6) and (2.7) that $\Theta_{ab} = u^\top \sin(\frac{\pi}{2}T)v$ while $\widetilde{\Theta}_{ab} = u^\top \times \sin(\frac{\pi}{2}\widehat{T})v$. Now, taking a linear approximation to $\sin(\cdot)$, we can write

$$\begin{aligned}
&\widetilde{\Theta}_{ab} - \Theta_{ab} \\
&\quad \approx \left\langle uv^\top \circ \frac{\pi}{2}\cos\left(\frac{\pi}{2}\widehat{T}\right), \widehat{T} - T \right\rangle \\
&\quad = \frac{1}{\binom{n}{2}} \sum_{i<i'} \left\langle uv^\top \circ \frac{\pi}{2}\cos\left(\frac{\pi}{2}\widehat{T}\right), \operatorname{sign}(X_i - X_{i'})\operatorname{sign}(X_i - X_{i'})^\top - T \right\rangle.
\end{aligned}$$

To study the variability of this error, we consider the kernel

$$g(X, X') = \operatorname{sign}(X - X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \operatorname{sign}(X - X').$$

We will see later on that understanding the behavior of this kernel will allow us to characterize the distribution of the oracle estimator $\widetilde{\Theta}_{ab}$, and from there, our empirical estimator $\breve{\Theta}_{ab}$ and ultimately $\breve{\Omega}_{ab}$. Of course, $g(X, X')$ itself depends on unknown quantities, namely u, v and T, so we replace these with their estimates in our empirical version of the kernel: define the (random) kernel

$$\breve{g}(X, X') = \operatorname{sign}(X - X')^\top \left(\breve{u}\breve{v}^\top \circ \cos\left(\frac{\pi}{2}\widehat{T}\right)\right) \operatorname{sign}(X - X'),$$

where

$$\breve{u}_a = 1, \qquad \breve{u}_b = 0, \qquad \breve{u}_I = -\breve{\gamma}_a \quad \text{and} \quad \breve{v}_a = 0, \qquad \breve{v}_b = 1, \qquad \breve{v}_I = -\breve{\gamma}_b.$$

(Note that we have defined $\breve{u}$ and $\breve{v}$ so that $\breve{\Theta}_{ab} = \breve{u}^\top \widehat{\Sigma}\breve{v}$.) Then define

$$\breve{S}_{ab} = \frac{\pi}{\det(\breve{\Theta})} \cdot \sqrt{\frac{1}{n}\sum_i \left(\frac{1}{n-1}\sum_{i' \neq i} \breve{g}(X_i, X_{i'}) - \mathsf{mean}(\breve{g})\right)^2},$$

where $\mathsf{mean}(\breve{g}) = \binom{n}{2}^{-1} \sum_{i<i'} \breve{g}(X_i, X_{i'})$. We will see later on that $\breve{S}_{ab}^2/n$ estimates the variance of $\breve{\Theta}_{ab}$ and that the expression above arises naturally from the theory of U-statistics.

Our main result, Theorem 3.5 below, will prove that $\sqrt{n} \cdot \frac{\breve{\Omega}_{ab} - \Omega_{ab}}{\breve{S}_{ab}}$ follows a distribution that is approximately standard normal. Therefore, an approximate $(1 - \alpha)$-confidence interval for Ω_{ab} is given by

$$\breve{\Omega}_{ab} \pm z_{\alpha/2} \cdot \frac{\breve{S}_{ab}}{\sqrt{n}}, \tag{2.10}$$

where $z_{\alpha/2}$ is the appropriate quantile of the normal distribution, that is, $\mathbb{P}\{N(0, 1) > z_{\alpha/2}\} = \alpha/2$.

Notation for fixed versus random quantities. From this point on, as much as possible throughout the main body of the paper, quantities that depend on the data and depend on the initial estimates $\check{\gamma}_a$, $\check{\gamma}_b$ are denoted with a "check" accent, for example, $\check{\Theta}$. Quantities that depend on the data, but do not depend on $\check{\gamma}_a$, $\check{\gamma}_b$, are denoted with a "hat" accent, for example, $\widehat{\Sigma}$. Any quantities with neither a "hat" nor a "check" are population quantities, that is, they are not random. Two important exceptions are the data itself, $X_1, \ldots, X_n$, and the oracle estimator, $\widetilde{\Theta}$, which is of course data-dependent (but does not depend on $\check{\gamma}_a$, $\check{\gamma}_b$).

3. Main results. In this section, we give a theoretical result showing that the confidence interval constructed in (2.10) has asymptotically the correct coverage probability, as long as we have reasonably accurate estimators of $\gamma_a = \Sigma_I^{-1}\Sigma_{Ia}$ and $\gamma_b = \Sigma_I^{-1}\Sigma_{Ib}$. Our asymptotic result considers a problem whose dimension $p_n \geqslant 2$ grows with the sample size n. We also allow for the sparsity level in the true inverse covariance matrix $\Omega \in \mathbb{R}^{p_n \times p_n}$ to grow.[2] We use k_n to denote an approximate bound on the sparsity in each column of Ω (details given below).

We begin by stating several assumptions on the distribution of the data and on the initial estimators $\check{\gamma}_a$ and $\check{\gamma}_b$. All of the constants appearing in these assumptions should be interpreted as values that do not depend on the dimensions (n, p_n, k_n) of the problem.

ASSUMPTION 3.1. The data points $X_1, \ldots, X_n \in \mathbb{R}^{p_n}$ are i.i.d. draws from a transelliptical distribution, $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_{p_n})$, where $f_1, \ldots, f_{p_n}$ are any strictly monotone functions, $\xi \geqslant 0$ is any random variable with $\mathbb{P}\{\xi = 0\} = 0$, and the covariance matrix $\Sigma \in \mathbb{R}^{p_n \times p_n}$ is positive definite, with $\operatorname{diag}(\Sigma) = \mathbf{1}$ and bounded condition number,

$$\mathsf{C}(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leqslant C_{\mathsf{cov}},$$

for some constant C_{cov}.

ASSUMPTION 3.2. The ath and bth columns of the true inverse covariance Ω, denoted by Ω_a and Ω_b, are approximately k_n-sparse, with

$$\|\Omega_a\|_1 \vee \|\Omega_b\|_1 \leqslant C_{\mathsf{sparse}}\sqrt{k_n},$$

for some constant C_{sparse}.

[2]While Σ, Ω, etc. all depend on the sample size n since the dimension of the problem grows, we abuse notation and do not write Σ_n, Ω_n, etc.; the dependence on n is implicit.

ASSUMPTION 3.3. For some constant C_{est} and for some $\delta_n > 0$, with probability at least $1 - \delta_n$, for each $c = a, b$, the preliminary estimate $\breve{\gamma}_c$ of the vector γ_c satisfies

$$\|\breve{\gamma}_c - \gamma_c\|_2 \leqslant C_{\mathsf{est}}\sqrt{\frac{k_n \log(p_n)}{n}}, \qquad \|\breve{\gamma}_c - \gamma_c\|_1 \leqslant C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}}. \tag{3.1}$$

ASSUMPTION 3.4. Define the kernel $h(X, X') = \operatorname{sign}(X - X') \otimes \operatorname{sign}(X - X') \in \mathbb{R}^{p_n^2}$ and let $h_1(X) = \mathbb{E}[h(X, X') \mid X]$. Define the total variance $\Sigma_h = \mathsf{Var}(h(X, X'))$ and the conditional $\Sigma_{h_1} = \mathsf{Var}(h_1(X))$, where $X, X' \overset{\text{i.i.d.}}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_{p_n})$. Then for some constant $C_{\mathsf{kernel}} > 0$,[3]

$$C_{\mathsf{kernel}} \cdot \Sigma_h \preceq \Sigma_{h_1} \preceq \Sigma_h.$$

Assumption 3.1 assumes that the smallest and largest eigenvalues of the correlation matrix Σ are bounded away from zero and infinity, respectively. This assumption is commonly assumed in the literature on learning structure of probabilistic graphical models [Ravikumar et al. (2011), Liu, Lafferty and Wasserman (2009), Liu et al. (2012)]. Assumption 3.2 does not require that the precision matrix Ω be exactly sparse, which is commonly assumed in the literature on exact graph recovery [see, e.g., Ravikumar et al. (2011)], but only requires that rows Ω_a and Ω_b have an ℓ_1 norm that does not grow too fast. Note that if Ω_c, for $c = a, b$, is k_n-sparse vector, then

$$\|\Omega_c\|_1 \leqslant \sqrt{k_n}\|\Omega_c\|_2 \leqslant \sqrt{k_n}\lambda_{\max}(\Omega) \leqslant C_{\mathsf{cov}}\sqrt{k_n}$$

and we could then set $C_{\mathsf{sparse}} = C_{\mathsf{cov}}$. Assumption 3.3 is a high-level condition, which assumes existence of initial estimators of γ_a and γ_b that converge at a fast enough rate. In the next section, we will see that Assumption 3.1 together with a stronger version of Assumption 3.2 are sufficient for Assumption 3.3 to be satisfied with a specific estimator that is efficient to compute. Finally, Assumption 3.4 is imposed to allow for estimation of the asymptotic variance $\breve{\Omega}_{ab}$. While Assumption 3.1 depends only on the correlation matrix Σ without reference to the distribution of the radius ξ, Assumption 3.4 depends on both Σ and ξ and, therefore, cannot be derived as a consequence of the choice of Σ.

We now state our main result.

THEOREM 3.5. *Under Assumptions 3.1, 3.2, 3.2 and 3.4, there exists a constant C_{converge}, depending on $C_{\mathsf{cov}}, C_{\mathsf{sparse}}, C_{\mathsf{est}}, C_{\mathsf{kernel}}$ but not on the dimensions*

[3]Here, we use the positive semidefinite ordering on matrices, that is, $A \succeq B$ if $A - B \succeq 0$. Note that the second part of the inequality, $\Sigma_{h_1} \preceq \Sigma_h$, is always true by the law of total variance.

(n, p_n, k_n) *of the problem, such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\sqrt{n}\cdot\frac{\breve{\Omega}_{ab}-\Omega_{ab}}{\breve{S}_{ab}}\leqslant t\right\}-\Phi(t)\right|\leqslant C_{\mathsf{converge}}\cdot\sqrt{\frac{k_n^2\log^2(p_n)}{n}}+\frac{1}{p_n}+\delta_n.$$

We note that the result holds uniformly over a large class of data generating processes satisfying Assumptions 3.1, 3.2, 3.3 and 3.4, which are relatively weak assumptions compared to much of the sparse estimation and inference literature; we emphasize that the result holds without requiring exact model selection or oracle properties, which hold only for restrictive sequences of data generating processes. For example, we do not require the "beta-min" condition (i.e., a lower bound on $|\Omega_{ab}|$ for all true edges) or any incoherence conditions [Bühlmann and van de Geer (2011)], which may be implausible in practice. Instead of requiring perfect model selection, we only require estimation consistency as given in Assumption 3.3; our weaker assumptions would not be sufficient to guarantee model selection consistency.

As an immediate corollary, we see that the confidence interval constructed in (2.10) is asymptotically correct.

COROLLARY 3.6. *Under the assumptions and notation of Theorem 3.5, the* $(1-\alpha)$*-confidence interval constructed in* (2.10) *fails to cover the true parameter* Ω_{ab} *with probability no higher than*

$$\alpha+2\left[C_{\mathsf{converge}}\cdot\sqrt{\frac{k_n^2\log^2(p_n)}{n}}+\frac{1}{p_n}+\delta_n\right].$$

Again this result holds uniformly over a large class of data generating distributions.

Theorem 3.5 is striking as it shows that we can form an asymptotically normal estimator of Ω_{ab} under the transelliptical distribution family with the sample complexity $n=\Omega(k_n^2\log^2(p_n))$. This sample size requirement was shown to be optimal for obtaining an asymptotically normal estimator of an element in a precision matrix from multivariate normal data [Ren et al. (2015)]. More precisely, let[4]

$$\mathcal{G}_0(c_0,c_1,k_n)=\Bigg\{\Omega=(\Omega_{ab})_{a,b\in[p_n]}:\max_{a\in[p_n]}\sum_{b\neq a}\mathbb{1}\{\Omega_{ab}\neq 0\}\leqslant k_n,$$
$$\text{and } c_0\leqslant\lambda_{\min}(\Omega)\leqslant\lambda_{\max}(\Omega)\leqslant c_1\Bigg\},$$

[4]In their work, the constants c_0, c_1 are instead denoted by a constant $M\geqslant c_0^{-1}\vee c_1$; we use different notation here to distinguish from the M used in Gu et al. (2015) which plays a very different role, and which we denote by M_n as it is not necessarily constant.

where $c_0, c_1 > 0$ are constants. Then Theorem 1 in Ren et al. (2015) proves

$$\inf_{a,b} \inf_{\breve{\Omega}_{ab}} \sup_{\mathcal{G}_0(c_0,c_1,k_n)} \mathbb{P}\big\{|\breve{\Omega}_{ab} - \Omega_{ab}| \geqslant \epsilon_0\big(n^{-1}k_n \log(p_n) \vee n^{-1/2}\big)\big\} \geqslant \epsilon_0$$

and, therefore, our estimator is rate optimal in terms of the sample size scaling. (Above, the infimum is taken over any estimator $\breve{\Omega}_{ab}$ which is a measurable function of the data.) We can also consider a related optimality question: whether the confidence interval we produce has the optimal (i.e., lowest possible) width, given the desired coverage level. In the Gaussian setting, Ren et al.'s (2015) method produces an interval which has asymptotically minimal length at the given sample size, due to the fact that the variance of their estimator matches the Fisher information. Our ROCKET method does not enjoy this theoretical property, but empirically we observe that our confidence intervals are only slightly wider than those produced by Ren et al.'s (2015) method, for Gaussian data.

At this point, it is also worth mentioning the result of Gu et al. (2015), who study inference under Gaussian copula graphical models. They base their inference procedure on decorrelating a pseudo score function for the parameter of interest and showing that it is normally distributed. Their main result, stated in Theorem 4.10, requires the sample size to satisfy

$$k_n^3 M_n^6 \left(\frac{\log(p_n)}{n}\right)^{3/2} + k_n^2 M_n^3 \frac{\log(p_n)}{n} = o\big(n^{-1/2}\big),$$

where $M_n = \max_{a\in[p_n]} \sum_{b\in[p_n]} |\Omega_{ab}|$. As M_n can be potentially as large as $\sqrt{k_n}$, it is immediately clear that our result achieves much better scaling on the sample size.

3.1. *Initial estimators.* The validity of our inference method relies in part on the accuracy of the initial estimators $\breve{\gamma}_a$ and $\breve{\gamma}_b$, which are assumed to satisfy error bounds with high probability as stated in Assumption 3.3—that is, with high probability, we have

$$\|\breve{\gamma}_c - \gamma_c\|_2 \leqslant C_{\mathsf{est}} \sqrt{\frac{k_n \log(p_n)}{n}}, \qquad \|\breve{\gamma}_c - \gamma_c\|_1 \leqslant C_{\mathsf{est}} \sqrt{\frac{k_n^2 \log(p_n)}{n}},$$

for $c = a, b$, where C_{est} is some constant. Below, we will prove that these required error rates can be obtained, under an additional sparsity assumption, by the Lasso estimators

$$\breve{\gamma}_c = \operatorname*{argmin}_{\gamma\in\mathbb{R}^I;\, \|\gamma\|_1 \leqslant C_{\mathsf{cov}}\sqrt{2k_n}} \left\{ \frac{1}{2}\gamma^\top \widehat{\Sigma}_I \gamma - \gamma^\top \widehat{\Sigma}_{Ic} + \lambda \|\gamma\|_1 \right\} \tag{3.2}$$

for each $c = a, b$, when the penalty parameter λ is chosen appropriately. In fact, these optimization problems may not be convex, because $\widehat{\Sigma}_I$ will not necessarily be positive semidefinite.

We now turn to proving that any local minima for (3.2) for $c = a, b$ will satisfy the required error rates of Assumption 3.3. To proceed, we will use the theoretical results of Loh and Wainwright (2015), which gives a theory for local minimizers of nonconvex regularized objective functions. In particular, any local minimizers of the two optimization problems will satisfy requirements of Assumption 3.3 and, therefore, we only need to be able to run optimization algorithms that find local minima. We specialize their main result to our setting.

THEOREM 3.7 [Adapted from Loh and Wainwright (2015), Theorem 1]. *Consider any $n, p \geqslant 1$, any $A \in \mathbb{R}^{p\times p}$ and $z \in \mathbb{R}^p$, and any k-sparse $x^\star \in \mathbb{R}^p$ with $\|x^\star\|_1 \leqslant R$. Suppose that A satisfies restricted strong convexity conditions*

$$v^\top A v \geqslant \alpha_1 \|v\|_2^2 - \tau_1 \|v\|_1^2 \cdot \frac{\log(p)}{n}. \tag{3.3}$$

If

$$n \geqslant \frac{16R^2\tau_1 \max\{\alpha_1, \tau_1\}\log(p)}{\alpha_1^2} \tag{3.4}$$

and

$$\max\left\{4\|Ax^\star - z\|_\infty, 4\alpha_1\sqrt{\frac{\log(p)}{n}}\right\} \leqslant \lambda \leqslant \frac{\alpha_1}{6R} \tag{3.5}$$

then for any $\check{x}$ that is a local minimum of the objective function $\frac{1}{2}x^\top Ax - x^\top z + \lambda\|x\|_1$ over the set $\{x \in \mathbb{R}^d : \|x\|_1 \leqslant R\}$, it holds that

$$\|\check{x} - x^\star\|_2 \leqslant \frac{1.5\lambda\sqrt{k}}{\alpha_1} \quad \text{and} \quad \|\check{x} - x^\star\|_1 \leqslant \frac{6\lambda k}{\alpha_1}.$$

We apply Loh and Wainwright's (2015) results, Theorem 3.7, to our problem of estimating γ_a and γ_b in a setting where we assume exact sparsity. (It is likely that similar results would hold for approximate sparsity, but here we use exact sparsity to fit the assumptions of this existing theorem.)

COROLLARY 3.8. *Suppose that Assumption 3.1 holds. Assume additionally that the columns Ω_a, Ω_b of the true inverse covariance $\Omega = \Sigma^{-1}$ are k_n-sparse. Then there exist constants C_{sample}, C_{Lasso}, depending on C_{cov} but not on (n, k_n, p_n), such that if $n \geqslant C_{\mathsf{sample}} k_n \log(p_n)$ then, with probability at least $1 - \frac{1}{2p_n}$, any local minimizer $\check{\gamma}_a$ of the objective function*

$$\frac{1}{2}\gamma^\top \hat{\Sigma}_I \gamma - \gamma^\top \hat{\Sigma}_{Ia} + \lambda\|\gamma\|_1$$

over the set $\{\gamma \in \mathbb{R}^I : \|\gamma\|_1 \leqslant C_{\mathsf{cov}}\sqrt{2k_n}\}$ satisfies

$$\|\check{\gamma}_a - \gamma_a\|_2 \leqslant 3\sqrt{2}C_{\mathsf{cov}}\lambda\sqrt{k_n} \quad \text{and} \quad \|\check{\gamma}_a - \gamma_a\|_1 \leqslant 24C_{\mathsf{cov}}\lambda\sqrt{k_n},$$

where we choose $\lambda = C_{\mathsf{Lasso}} \cdot \sqrt{\frac{\log(p_n)}{n}}$. The same holds for estimating γ_b.

Using this corollary, we see that a local minimizer of (3.2), $\check{\gamma}_c$, satisfies Assumption 3.3 with $\delta_n = \frac{1}{p_n}$ and $C_{\mathsf{est}} = 24 C_{\mathsf{cov}} C_{\mathsf{Lasso}}$. We remark that in practice, the constant C_{Lasso} suggested by the theory is in general unknown, but choosing λ to be a small multiple of $\sqrt{\log(p_n)/n}$ generally performs well—for instance, $\lambda = 2.1\sqrt{\log(p_n)/n}$ as we use in our simulations, where the choice of the constant 2.1 ensures that the penalty term dominates the variance of the elements of the objective function's derivative, that is, the elements of $\widehat{\Sigma}_I \gamma - \widehat{\Sigma}_{Ic}$, at the true solution $\gamma = \gamma_c$.

To prove that this corollary follows from Loh and Wainwright's (2015) result (Theorem 3.7), it is sufficient to check that the restricted strong convexity condition (3.3) holds with high probability for the matrix $\widehat{\Sigma}_I$, and then compute the necessary values for λ and the other parameters of Theorem 3.7. The proof is technical and relies on novel results on concentration of the Kendall's tau correlation matrix (see the Supplementary Materials [Barber and Kolar (2018)]).

We have provided sufficient condition for a local minimizer of (3.2) to satisfy Assumption 3.3; however, many other estimators can be used as initial estimators. For example, one could use the Dantzig selector [Candes and Tao (2007)]. Potential benefits of the Dantzig selector over the optimization program in (3.2) are twofold. First, the optimization program is convex even when $\widehat{\Sigma}_I$ is not positive semidefinite. Second, one does not need to know an upper bound R on the ℓ_1 norm of Ω_c for $c = a, b$. Using the techniques similar to those used to prove Corollary 3.8, we can also prove that Assumption 3.3 holds when the Dantzig selector is used as an initial estimator. For large problems, however, Dantzig selector type methods are computationally much slower than Lasso type methods; in our empirical results, we implement the Lasso rather than the Dantzig selector since we study graphs with as many as 1000 nodes.

In practice, we have found that in simulations, using the Lasso for model selection, and then refitting without a penalty, leads to better empirical performance. Specifically, for each $c = a, b$, we first fit

$$\check{\gamma}_c^{\mathsf{Lasso}} = \underset{\gamma \in \mathbb{R}^I}{\operatorname{argmin}} \left\{ \frac{1}{2} \gamma^\top \widehat{\Sigma}_I \gamma - \gamma^\top \widehat{\Sigma}_{Ia} + \lambda \|\gamma\|_1 \right\};$$

or, more precisely, find a local minimum of this nonconvex optimization problem over the ball $\{\gamma : \|\gamma\|_1 \leqslant R\}$ for a large radius R. (In practice, every iteration will stay inside this ball; therefore, as long as we see convergence in our iterative algorithm for solving this nonconvex Lasso, we do not concern ourselves with this theoretical boundedness constraint.)

We then extract the combined support of the two solutions, $\check{J} = \mathsf{supp}(\check{\gamma}_a^{\mathsf{Lasso}}) \cup \mathsf{supp}(\check{\gamma}_b^{\mathsf{Lasso}})$, and refit the coefficients using least-squares:

$$\check{\gamma}_c = (\widehat{\Sigma}_{\check{J}})^{-1} \widehat{\Sigma}_{\check{J}c} \qquad \text{for } c = a, b.$$

Following the work of Belloni and Chernozhukov (2013) or Sun and Zhang (2012b), it can be shown that the refitted estimators also satisfy the Assumption 3.3; in practice, refitting improves the accuracy of these preliminary estimators by reducing shrinkage bias.

Finally, we remark that if we would like to perform inference for all $\binom{p_n}{2}$ potential edges, then we require $2 \cdot \binom{p_n}{2} \sim p_n^2$ many initial estimators to be computed; this is of course quite computationally demanding. However, Ren et al. (2015) propose a simple modification that significantly reduces computation time: for each node a we can first regress X_a on all the other variables; call this solution $\check{\gamma}_a^{\text{all}}$. Next, for any $b \neq a$, if $(\check{\gamma}_a^{\text{all}})_b = 0$, then this solution $\check{\gamma}_a^{\text{all}}$ is already optimal for regressing node a on nodes $I = [p_n] \backslash \{a, b\}$; this will be the case for most nodes b due to sparsity. With this modification, the actual number of regressions required is far smaller—if each node a forms edges with at most k_n other nodes (i.e., $\check{\gamma}_a^{\text{all}}$ is k_n-sparse), then we will require only $p_n(k_n + 1)$ many regressions in total to form all of the initial estimators.

4. Main technical tools. In this section, we outline the proof of Theorem 3.5 (Section 4.1) and state the key technical result that establishes the sign-sub-Gaussianity property of a vector X following a transelliptical distribution (Section 4.2). We also illustrate an application of this technical result to establishing a bound on $\hat{\Sigma} - \Sigma$ (Section 4.3).

4.1. *Sketch of proof for main result.* The proof of Theorem 3.5 has two key steps. First, in Step 1, we prove that the distribution of $\widetilde{\Theta}_{ab}$, the oracle estimator of Θ_{ab}, is asymptotically normal, with

$$\sqrt{n} \cdot \frac{\widetilde{\Theta}_{ab} - \Theta_{ab}}{S_{ab} \det(\Theta)} \to N(0, 1),$$

where S_{ab} is the asymptotic variance of $\tilde{\Omega}_{ab}$. (Explicit form of S_{ab} is given in the proof of Theorem 4.1.) Next, in Step 2, we prove that the difference between the estimator and the oracle estimator, $\check{\Theta} - \widetilde{\Theta}$, converges to zero at a fast rate and that the variance estimator $\check{S}_{ab}$ converges to S_{ab} at a fast rate. Combining these steps, we prove that $\check{\Omega}_{ab}$ is an asymptotically normal estimator of Ω_{ab}. The detailed proofs for each step are found in the Supplementary Materials [Barber and Kolar (2018)]. Here, we outline the main results for each step.

Step 1 establishes the Berry–Esseen-type bound for the centered and normalized oracle estimator $\sqrt{n} \cdot \frac{\widetilde{\Theta}_{ab} - \Theta_{ab}}{S_{ab} \cdot \det(\Theta)}$. We approximate the oracle estimator $\widetilde{\Theta}_{ab}$ by a linear function of the Kendall's tau statistic $\widehat{T}$, which is a U-statistic of the data. We prove that the variance of the linear approximation is bounded away from zero and apply existing results on convergence of U-statistics. The following result is proved in the Supplementary Materials [Barber and Kolar (2018)].

THEOREM 4.1. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Then there exist constants* C_{normal}, C_{variance} *depending on* C_{cov}, C_{sparse}, C_{kernel} *but not on* (n, p_n, k_n), *such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\sqrt{n}\cdot\frac{\widetilde{\Theta}_{ab}-\Theta_{ab}}{S_{ab}\cdot\det(\Theta)}\leqslant t\right\}-\Phi(t)\right|\leqslant C_{\mathsf{normal}}\cdot\frac{k_n\log(p_n)}{\sqrt{n}}+\frac{1}{2p_n},$$

where S_{ab} *is defined in the proof and satisfies* $S_{ab}\cdot\det(\Theta)\geqslant C_{\mathsf{variance}}>0$.

Step 2 contains the main challenge of this problem, since it requires strong results on the concentration properties of the Kendall's tau estimator $\widehat{\Sigma}$ of the covariance matrix Σ. The main ingredient for this step is a new result on "sign-sub-Gaussianity," that is, proving that the signs vector $\operatorname{sign}(X_i - X_{i'})$ is sub-Gaussian for i.i.d. observations X_i, $X_{i'}$. Our results on sign-sub-Gaussianity are discussed in Section 4.2 and their application to concentration of $\hat{\Sigma}$ around Σ is given in Section 4.3. Using these tools, we are able to prove the following theorem (proved in the Supplementary Materials [Barber and Kolar (2018)]).

THEOREM 4.2. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold. Then there exists a constant* C_{oracle}, *depending on* C_{cov}, C_{sparse}, C_{est} *but not on* (n, p_n, k_n), *such that, if*[5] $n\geqslant 15k_n\log(p_n)$, *then with probability at least* $1-\frac{1}{2p_n}-\delta_n$,

$$\|\breve{\Theta}-\widetilde{\Theta}\|_\infty\leqslant C_{\mathsf{oracle}}\cdot\frac{k_n\log(p_n)}{n}$$

and

$$\left|\breve{S}_{ab}\cdot\det(\breve{\Theta})-S_{ab}\cdot\det(\Theta)\right|\leqslant C_{\mathsf{oracle}}\cdot\sqrt{\frac{k_n^2\log(p_n)}{n}}.$$

4.2. *Sign-sub-Gaussian random vectors*. Recall the definition of a sub-Gaussian random vector.

DEFINITION 4.3. A random vector $X\in\mathbb{R}^p$ is C-sub-Gaussian if, for any fixed vector $v\in\mathbb{R}^p$, it holds that $\mathbb{E}[e^{v^\top X}]\leqslant e^{C\cdot\|v\|_2^2/2}$.

For graphical models where the data points X_i come from a sub-Gaussian distribution, the sample covariance matrix $\frac{1}{n}\sum_i(X_i-\overline{X})(X_i-\overline{X})^\top$, with $\overline{X}=\frac{1}{n}\sum_i X_i$, is known to concentrate near the population covariance, as measured by different norms. For example, elementwise convergence of the sample covariance to the population covariance, that is, convergence in $\|\cdot\|_\infty$, is sufficient to establish

[5]Note that the additional condition $n\geqslant 15k_n\log(p_n)$ can be assumed to hold in our main result Theorem 3.5, since if this inequality does not hold, then the claim in Theorem 3.5 is trivial.

rates of convergence for the graphical Lasso, CLIME or graphical Dantzig selector for estimating the sparse inverse covariance [Ravikumar et al. (2011), Cai, Liu and Luo (2011), Yuan (2010)]. Similar results can be obtained also for the transelliptical family, since $\|\hat{T} - T\|_\infty \leqslant C\sqrt{\log(p)/n}$, and hence $\|\hat{\Sigma} - \Sigma\|_\infty \leqslant C\sqrt{\log(p)/n}$, as was shown in Liu et al. (2012) and Liu, Han and Zhang (2012). However, in order to construct asymptotically normal estimators for the elements of the precision matrix, stronger results are needed about the convergence of the sample covariance to the population covariance [Ren et al. (2015)]. In particular, a result on convergence in spectral norm, uniformly over all sparse submatrices, is required. One can relate the convergence in the elementwise ℓ_∞ norm to (sparse) spectral norm convergence, however, this would lead to suboptimal sample size. One way to obtain a tight bound on the (sparse) spectral norm convergence is by utilizing sub-Gaussianity of the data points X_i. This is exactly what we proceed to establish.

Recall from (2.3) the Kendall's tau estimator of the covariance,

$$\widehat{\Sigma} = \sin\left(\frac{\pi}{2}\widehat{T}\right) \qquad \text{where } \widehat{T} = \frac{1}{\binom{n}{2}} \sum_{i<i'} \operatorname{sign}(X_i - X_{i'}) \operatorname{sign}(X_i - X_{i'})^\top.$$

Therefore, it is crucial to determine whether the vector $\operatorname{sign}(X_i - X_{i'})$ is itself sub-Gaussian, at a scale that does not depend heavily on the ambient dimension p_n.[6] Using past results on elliptical distributions, we can reduce to a simpler case using the arguments of Lindskog, McNeil and Schmock (2003) (proved in the Supplementary Materials [Barber and Kolar (2018)]).

LEMMA 4.4. *Let* $X, X' \overset{\text{i.i.d.}}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_p)$. *Suppose that* Σ *is positive definite, and that* $\xi > 0$ *with probability* 1. *Then* $\operatorname{sign}(X - X')$ *is equal in distribution to* $\operatorname{sign}(Z)$, *where* $Z \sim N(0, \Sigma)$.

Previous work has shown that a Gaussian random vector $Z \sim N(0, \Sigma)$ is "sign-sub-Gaussian," that is, $\operatorname{sign}(Z)$ is sub-Gaussian with variance proxy that depends on p_n only through $C(\Sigma)$, for special cases when the covariance Σ is identity or equicorrelation matrix [Han and Liu (2013)]. However, a result for general covariance structures was previously unknown.

In the following lemma, we resolve this question, proving that Gaussian vectors are sign-sub-Gaussian [recall $\mathsf{C}(\Sigma)$ is the condition number of Σ].

LEMMA 4.5. *Let* $Z \sim N(\mu, \Sigma)$. *Then* $\operatorname{sign}(Z) - \mathbb{E}[\operatorname{sign}(Z)]$ *is* $\mathsf{C}(\Sigma)$*-sub-Gaussian.*

[6]Note that $v^\top \operatorname{sign}(X_i - X_{i'})$ is obviously sub-Gaussian for any distribution on X, as it is a sum of sub-Gaussian random variables [since $\operatorname{sign}(\cdot)$ is bounded]; however, its scale could grow linearly with p_n.

This lemma is the primary tool for our main results in this paper—specifically, it is the key ingredient to the proof of Theorem 4.2, which bounds the errors $\breve{\Theta} - \widetilde{\Theta}$ and $\breve{S}_{ab} \cdot \det(\breve{\Theta}) - S_{ab} \cdot \det(\Theta)$. Lemma 4.5 is proved in the Supplementary Materials [Barber and Kolar (2018)]. We also use this result in establishing results in the following section.

4.3. *Deterministic and probabilistic bounds on* $\widehat{\Sigma} - \Sigma$. Lemma 4.5 is instrumental in obtaining probabilistic bounds on $\widehat{\Sigma} - \Sigma$. Results given in this section are crucial for establishing Theorem 4.2 and Corollary 3.8.

Let $\mathcal{S}_k$ be the set of k-sparse vectors in the unit ball,

$$\mathcal{S}_k = \{u \in \mathbb{R}^p : \|u\|_2 \leqslant 1, \|u\|_0 \leqslant k\},$$

and abusing notation, let $\|\cdot\|_{\mathcal{S}_k}$ denote the sparse spectral norm for matrices, that is, $\|M\|_{\mathcal{S}_k} = \max_{u,v \in \mathcal{S}_k} u^\top M v$.

The following lemma provides a bound on the error in Kendall's tau, that is, on $\widehat{T} - T$, in this sparse spectral norm (with the proof given in the Supplementary Materials [Barber and Kolar (2018)]).

LEMMA 4.6. *Suppose that* $k \geqslant 1$ *and* $\delta \in (0,1)$ *satisfy* $\log(2/\delta) + 2k \times \log(12p) \leqslant n$. *Then with probability at least* $1 - \delta$ *it holds that*

$$\|\widehat{T} - T\|_{\mathcal{S}_k} \leqslant 32(1+\sqrt{5})\mathsf{C}(\Sigma) \cdot \sqrt{\frac{\log(2/\delta) + 2k\log(12p)}{n}}.$$

Next, we relate $\hat{\Sigma}$ to $\hat{T}$, with the following deterministic bound on the sparse spectral norm of the error of the covariance estimator $\widehat{\Sigma}$, which is proven in the Supplementary Materials [Barber and Kolar (2018)].

LEMMA 4.7. *The following bound holds deterministically*: *for any* $k \geqslant 1$,

$$\|\hat{\Sigma} - \Sigma\|_{\mathcal{S}_k} \leqslant \frac{\pi^2}{8} \cdot k \|\widehat{T} - T\|_\infty^2 + 2\pi \|\widehat{T} - T\|_{\mathcal{S}_k}. \tag{4.1}$$

A result in de la Peña and Giné (1999,Theorem 4.1.8) bounds $\|\widehat{T} - T\|_\infty$ with high probability (details of this bound are given in the Supplementary Materials [Barber and Kolar (2018)]). Combining the bound on $\|\widehat{T} - T\|_\infty$ with Lemmas 4.6 and 4.7, we immediately obtain the following corollary.

COROLLARY 4.8. *Take any* $\delta_1, \delta_2 \in (0,1)$ *and any* $k \geqslant 1$ *such that* $\log(2/\delta_2) + 2k\log(12p) \leqslant n$. *Then, with probability at least* $1 - \delta_1 - \delta_2$, *the following bound on* $\widehat{\Sigma} - \Sigma$ *holds*:

$$\begin{aligned}\|\hat{\Sigma} - \Sigma\|_{\mathcal{S}_k} \leqslant{}& \frac{\pi^2}{8} \cdot k \cdot \frac{4\log(2\binom{p}{2}/\delta_1)}{n} \\ &+ 2\pi \cdot 32(1+\sqrt{5})\mathsf{C}(\Sigma) \cdot \sqrt{\frac{\log(2/\delta_2) + 2k\log(12p)}{n}}.\end{aligned} \tag{4.2}$$

Finally, we use a result based on the work of Sun and Zhang (2012b), in order to extend this sparse spectral norm bound to a bound holding for all approximately sparse vectors u and v.

LEMMA 4.9 [Based on Proposition 5 of Sun and Zhang (2012b)]. *For any fixed matrix $M \in \mathbb{R}^{p\times p}$ and vectors $u, v \in \mathbb{R}^p$, and any $k \geqslant 1$,*

$$\left|u^\top M v\right| \leqslant \left(\|u\|_2 + \|u\|_1/\sqrt{k}\right) \cdot \left(\|v\|_2 + \|v\|_1/\sqrt{k}\right) \cdot \|M\|_{\mathcal{S}_k}.$$

Results of Lemma 4.6 and Corollary 4.8 can be compared to Theorem 2 in Mitra and Zhang (2014), which proves essentially the same result for the Kendall's tau estimate of Σ, but only for the nonparanormal (Gaussian copula) model; their technique does not extend immediately to the transelliptical model. When $C(\Sigma) = O(1)$, we extend their result to the transelliptical model and, as a special case, this provides an alternative proof for their result on the Gaussian copula model. We note that their result does not depend on the condition number of the covariance matrix, but only on the maximum eigenvalue. However, in the context of graphical models it is commonly assumed that the smallest eigenvalue is a constant. Furthermore, our results in Lemma 4.6 and Corollary 4.8 can also be compared with Theorem 4.10 of Han and Liu (2013), which give similar bounds on the spectral norm of sparse submatrices of $\widehat{T} - T$ and $\widehat{\Sigma} - \Sigma$, but with a sign-sub-Gaussianity assumption on the distribution. We rigorously establish the same bounds for all well-conditioned covariance matrices, without explicitly making the sign-sub-Gaussian assumption.

5. Simulation studies. In this section, we illustrate finite sample properties of ROCKET described in Section 2 on simulated data. (A real data experiment, and some additional simulations, are presented in the Supplementary Materials [Barber and Kolar (2018)].)

We use ROCKET to construct confidence intervals for edge parameters and report empirical coverage probabilities as well as the length of constructed intervals. For comparison, we also construct confidence intervals using the procedure of Ren et al. (2015), which is based on the Pearson correlation matrix, a nonparanormal estimator of the correlation matrix (NPN) proposed in Liu, Lafferty and Wasserman (2009), and the pseudo score procedure of Gu et al. (2015). For the first two methods, we use the plug-in estimate of the correlation matrix together with (2.8) to estimate Ω_{ab}. Recall that Liu, Lafferty and Wasserman (2009) estimate the correlation matrix based on the marginal transformation of the observed data. Let

$$\tilde{F}_a(x) = \begin{cases} \delta_n & \text{if } \hat{F}_a(x) < \delta_n, \\ \hat{F}_a(x) & \text{if } \delta_n \leqslant \hat{F}_a(x) \leqslant 1 - \delta_n, \\ 1 - \delta_n & \text{if } \hat{F}_a(x) > 1 - \delta_n, \end{cases}$$

where $\hat{F}_a(x) = n^{-1}\sum_i \mathbb{1}\{X_{ia} < x\}$ is the empirical CDF of X_a and $\delta_n = (4n^{1/4}\sqrt{\pi \log(n)})^{-1}$. The correlation matrix $\hat{\Sigma} = (\hat{\Sigma}_{ab})_{ab}$ is then estimated as $\hat{\Sigma}_{ab} = \widehat{\mathrm{Corr}}(\Phi(\tilde{F}_a(X_{ia})), \Phi(\tilde{F}_b(X_{ib})))$. Asymptotic variance of estimators of Ω_{ab} based on the Pearson or nonparanormal correlation matrix is obtained as $\breve{S}^2_{ab} = n^{-1}(\breve{\Omega}_{aa}\breve{\Omega}_{bb} + \breve{\Omega}^2_{ab})$. Gu et al. (2015) estimate Ω_{ab} as

$$\breve{\Omega}^{\mathrm{PS}}_{ab} = \frac{\hat{\Omega}_{ab}((\hat{\Omega}\hat{\Sigma})_{ab} + (\hat{\Sigma}\hat{\Omega})_{ab}) - (\hat{\Omega}\hat{\Sigma}\hat{\Omega})_{ab}}{(\hat{\Omega}\hat{\Sigma})_{ab} + (\hat{\Sigma}\hat{\Omega})_{ab} - 1},$$

where $\hat{\Sigma}$ is the Kendall's tau estimator of the covariance matrix in (2.3) and $\hat{\Omega}$ is an initial estimator of the precision matrix. Under suitable conditions, $\hat{\Omega}^{\mathrm{PS}}_{ab}$ is asymptotically normal with the asymptotic variance that can be consistently estimated as in Corollary 4.12 of Gu et al. (2015). Gu et al. (2015) suggest using the CLIME estimator [Cai, Liu and Luo (2011)] to construct $\hat{\Omega}$, however, we find that empirically the method performs better using lasso-with-refitting to estimate each row of Ω, similar to Sun and Zhang (2012b). For all simulations, we set the tuning parameter $\lambda = 2.1\sqrt{\log(p_n)/n}$, as suggested by our theory—this constant is large enough so that the penalty dominates the variance of each element of the score. All computations are carried out in Matlab.

Simulation 1. We generate data from the model $X \sim \mathsf{E}(0, \Sigma, \xi)$, where ξ follows a t-distribution with 5 degrees of freedom. The inverse covariance matrix Ω encodes a grid where each node is connected to its four nearest neighbors with the nonzero elements of Ω^0 equal to $\omega = 0.24$. Diagonal element of Ω^0 are equal to 1. Let $(\Omega^0)^{-1} = \Sigma^0$. Then set $\Sigma = (\mathrm{diag}(\Sigma^0))^{-1/2}\Sigma^0(\mathrm{diag}(\Sigma^0))^{-1/2}$ and $\Omega = \Sigma^{-1}$. (Additional simulations in the Supplementary Materials [Barber and Kolar (2018)] show the same experiment on a chain graph structure.)

We take a grid of size 30×30 (so that $p_n = 900$) and take sample size $n = 400$. Figure 2 shows quantile-quantile (Q-Q) plots based on 1000 independent realizations of the test statistic error, $\sqrt{n} \cdot \frac{\breve{\Omega}_{ab} - \Omega_{ab}}{\breve{S}_{ab}}$, for the four methods together with the reference line showing quantiles of the standard normal distribution. From this figure, we observe that the quantiles of the test statistic error $\sqrt{n}\frac{\breve{\Omega}_{ab} - \Omega_{ab}}{\breve{S}_{ab}}$ based on ROCKET is closest to the quantiles of the standard normal random variable. We further quantify these results in Table 1, which reports empirical coverage and width of the confidence intervals based on $\sqrt{n}\frac{\breve{\Omega}_{ab} - \Omega_{ab}}{\breve{S}_{ab}}$. From the table, we can observe that the coverage of the confidence intervals based on ROCKET and the pseudo score are closest to nominal coverage of 95%. The three node pairs displayed in this figure and table, namely $\omega_{(2,2),(2,3)}$, $\omega_{(2,2),(3,3)}$, $\omega_{(2,2),(10,10)}$, correspond to a true edge, a nonedge between nearby nodes that is therefore easy to mistake for an edge, and a nonedge between distant nodes, respectively.

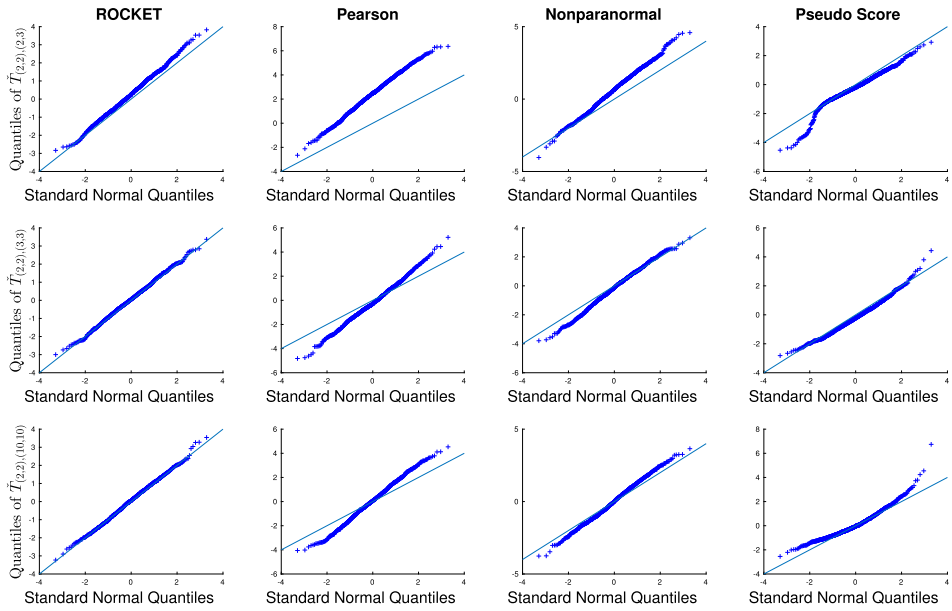

FIG. 2. *Simulation* 1 (*transelliptical data*). *Q-Q plot of* $\sqrt{n} \cdot \frac{\check{\Omega}_{ab} - \Omega_{ab}}{\check{S}_{ab}}$ *when* Ω *corresponds to a grid graph structure. Row* 1 *corresponds to an edge, row* 2 *to a close nonedge, and row* 3 *to a far nonedge.*

These results are not surprising, since neither the Pearson nor the nonparanormal correlation matrix consistently estimate the true Σ. In contrast, both ROCKET and the pseudo score method are able to construct a test statistic $\sqrt{n} \cdot \frac{\check{\Omega}_{ab}}{\check{S}_{ab}}$ that is asymptotically distributed as a normal random variable. The asymptotic distribution provides a good approximation to the finite sample distribution of $\sqrt{n} \cdot \frac{\check{\Omega}_{ab} - \Omega_{ab}}{\check{S}_{ab}}$.

Simulation 2. We illustrate performance of ROCKET when data are generated from a normal and nonparanormal distribution. We consider Ω correspond-

TABLE 1
Simulation 1 (*transelliptical data*). *Percent empirical coverage* (*average length*) *of* 95% *confidence intervals based on* 1000 *independent simulation runs*

	ROCKET	**Pearson**	**NPN**	**Pseudo score**
$\omega_{(2,2),(2,3)} = 0.37$	94.6 (0.51)	36.6 (0.88)	82.4 (0.48)	92.2 (0.52)
$\omega_{(2,2),(3,3)} = 0$	94.3 (0.53)	81.0 (0.86)	88.3 (0.47)	94.8 (0.50)
$\omega_{(2,2),(10,10)} = 0$	94.9 (0.56)	78.3 (0.88)	89.1 (0.48)	95.5 (0.53)

TABLE 2
Simulation 2 (*Gaussian and nonparanormal data*). *Percent empirical coverage* (*average length*) *of* 95% *confidence intervals based on* 1000 *independent simulation runs*

		ROCKET	**Pearson**	**NPN**	**Pseudo score**
Gaussian	$\omega_{(2,2),(2,3)} = 0.37$	95.6 (0.33)	94.1 (0.32)	94.4 (0.32)	95.9 (0.33)
	$\omega_{(2,2),(3,3)} = 0$	95.4 (0.35)	95.9 (0.34)	95.3 (0.34)	95.0 (0.35)
	$\omega_{(2,2),(10,10)} = 0$	96.0 (0.36)	95.3 (0.35)	95.6 (0.35)	94.8 (0.36)
Transf. Gaussian	$\omega_{(2,2),(2,3)} = 0.37$	95.1 (0.35)	12.0 (0.33)	94.9 (0.33)	95.3 (0.35)
	$\omega_{(2,2),(3,3)} = 0$	95.3 (0.35)	93.1 (0.32)	94.9 (0.34)	94.7 (0.34)
	$\omega_{(2,2),(10,10)} = 0$	93.7 (0.36)	93.2 (0.32)	94.2 (0.35)	94.8 (0.39)

ing to a grid as in Simulation 1, and generate $n = 400$ samples from $N(0, \Omega^{-1})$ and $\mathsf{NPN}(\Omega^{-1}; \tilde{f}_1, \ldots, \tilde{f}_p)$, where $\tilde{f}_j = f_{\mathrm{mod}(j-1,5)+1}$ with $f_1(x) = x$, $f_2(x) = \mathrm{sign}(x)\sqrt{|x|}$, $f_3(x) = x^3$, $f_4(x) = \Phi(x)$, $f_5(x) = \exp(x)$.

Table 2 summarizes results from the simulation. We observe that when data are multivariate normal all methods perform well, with ROCKET and the pseudo score having slightly wider intervals, but with similar coverage. When data are generated from a nonparanormal distribution, using the Pearson correlation in (2.8) results in confidence intervals that do not have nominal coverage due to the bias. In this setting, nonparanormal estimator, ROCKET and the pseudo score still have the correct nominal coverage. Note however that when Kendall's tau is equal to zero, Pearson correlation is also equal to zero, and coverage for Pearson improves. See, for example, coverage for $\omega_{(2,2),(3,3)}$ and $\omega_{(2,2),(10,10)}$.

Simulation 3. In this simulation, we illustrate the power of a test based on the statistic $\sqrt{n} \cdot \frac{\breve{\Omega}_{ab}}{\breve{S}_{ab}}$ to reject the null hypothesis $H_{0,ab} : \Omega_{ab} = 0$. Samples are generated from the $\mathsf{E}(0, \Sigma, \xi)$ with ξ having χ_{p_n}, t_5, and t_1 distribution and the covariance matrix is of the form $\Sigma = I_P + E$ where $E_{12} = E_{21} = \rho$ and all other entries zero, with $p_n = 1000$ and $n = 400$. Note that $\xi \sim \chi_{p_n}$ implies that X is multivariate normal. We also consider marginal transformation of X as described in Simulation 2. Figure 3 plots empirical power curves based on 1000 independent simulation runs for different settings. When the data follow a normal distribution all methods have similar power. For other distributions, tests based on Pearson and nonparanormal correlation do not have correct coverage and are shown for illustrative purpose only.

6. Discussion. We have proposed a novel procedure ROCKET for inference on elements of the latent inverse correlation matrix under high-dimensional elliptical copula models. Our paper has established a surprising result, which states that ROCKET produces an asymptotically normal estimator for an element of the

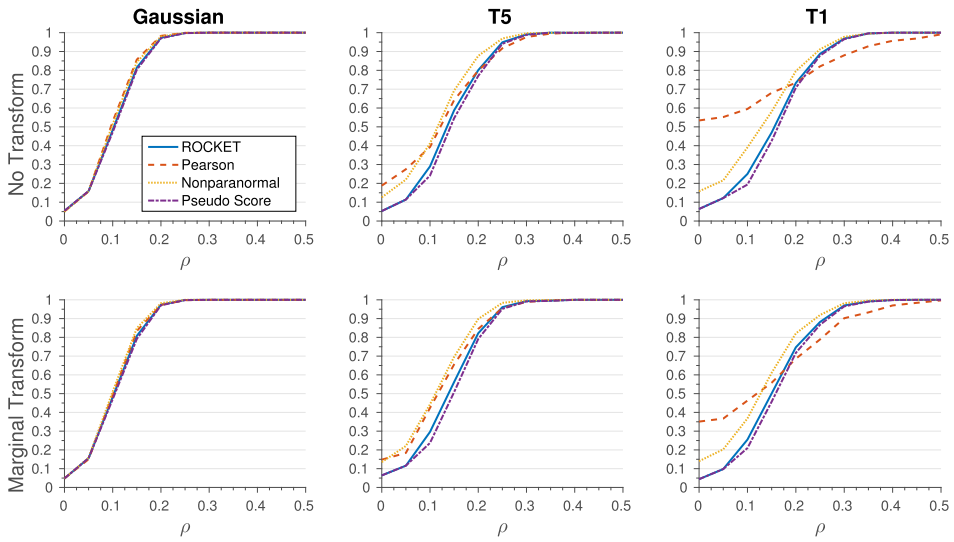

FIG. 3. *Simulation* 3. *Power plots for simulated data generated from a Gaussian distribution, and from a multivariate t distribution with* 5 *d.f. or with* 1 *d.f.*

inverse correlation matrix in an elliptical copula model with the same sample complexity that is required to obtain an asymptotically normal estimator for an element in the precision matrix under a multivariate normal distribution. Furthermore, this sample complexity is optimal [Ren et al. (2015)]. The result is surprising as the family of elliptical copula models is much larger than the family of multivariate normal distributions. For example, it contains distributions with heavy tail dependence as discussed in Section 2. ROCKET achieves the optimal requirement on the sample size without knowledge of the marginal transformation. Our result is also of significant practical importance. Since normal distribution is only a convenient mathematical approximation to data generating process, we recommend using ROCKET whenever making inference about inverse correlation matrix, instead of methods that heavily rely on normality. From simulation studies, even when data are generated from a normal distribution, ROCKET does not lose power compared to procedures that were specifically developed for inference under normality.

The main technical tool developed in the paper establishes that the sign of normal random vector, taken elementwise, is itself a sub-Gaussian random variable with the sub-Gaussian parameter depending on the condition number of the covariance matrix Σ (but not on the dimension p_n). Based on this result, we were able to establish a tight tail bound on the deviation of sparse eigenvalues of the Kendall's tau matrix $\hat{T}$. This result is of independent interest and it would allow us to improve a number of recent results on sparse principal component analysis, factor models and estimation of structured covariance matrices [Mitra and Zhang (2014), Han and Liu (2013), Fan, Han and Liu (2014)]. The sharpest result on the

nonparametric estimation of correlation matrices in spectral norm under a Gaussian copula model was established in [Mitra and Zhang (2014)]. Our results establish a similar result for the family of elliptical copula models and provide an alternative proof for the Gaussian copula model.

Acknowledgment. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

SUPPLEMENTARY MATERIAL

Supplement to "ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models" (DOI: 10.1214/17-AOS1663SUPP; .pdf). In the supplementary materials, we provide additional experimental results (as described in Section 5), as well as details for all proofs of the theoretical results provided in this paper.

REFERENCES

BARBER, R. F. and KOLAR, M. (2018). Supplement to "ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models." DOI:10.1214/17-AOS1663SUPP.

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2013a). Uniform post selection inference for LAD regression models. Preprint. Available at arXiv:1304.0282.

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2013b). Robust inference in high-dimensional approximately sparse quantile regression models. Preprint. Available at arXiv:1312.7186.

BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2013). Honest confidence regions for logistic regression with a large number of controls. Preprint. Available at arXiv:1304.3969.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CHEN, M., REN, Z., ZHAO, H. and ZHOU, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* **111** 394–406. MR3494667

CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. MR3640193

D'ASPREMONT, A., BANERJEE, O. and EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568

DE LA PEÑA, V. H. and GINÉ, E. (1999). *Decoupling*: *From Dependence to Independence*. Springer, New York. MR1666908

EMBRECHTS, P., LINDSKOG, F. and MCNEIL, A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance* (S. T. Rachev, ed.) 329–384. Elsevier, Amsterdam.

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FAN, J., HAN, F. and LIU, H. (2014). PAGE: Robust pattern guided estimation of large covariance matrix. Technical report, Princeton Univ., Princeton, NJ.

FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*. *Monographs on Statistics and Applied Probability* **36**. Chapman and Hall, Ltd., London. MR1071174

FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23. MR3397349

FRIEDMAN, J. H., HASTIE, T. J. and TIBSHIRANI, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GU, Q., CAO, Y., NING, Y. and LIU, H. (2015). Local and global inference for high dimensional Gaussian copula graphical models. Preprint. Available at arXiv:1502.02347.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011a). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011b). Asymptotic properties of the joint neighborhood selection method for estimating categorical Markov networks. Technical report, Univ. Michigan, Ann Arbor, MI.

HAN, F. and LIU, H. (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. Preprint. Available at arXiv:1305.6916.

HÖFLING, H. and TIBSHIRANI, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

JAVANMARD, A. and MONTANARI, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. Preprint. Available at arXiv:1311.0274.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

KLÜPPELBERG, C., KUHN, G. and PENG, L. (2008). Semi-parametric models for the multivariate tail dependence function—The asymptotically dependent case. *Scand. J. Stat.* **35** 701–718. MR2468871

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford Univ. Press, New York. MR1419991

LEE, J. D. and HASTIE, T. J. (2012). Learning mixed graphical models. Preprint. Available at arXiv:1205.5012.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

LINDSKOG, F., MCNEIL, A. and SCHMOCK, U. (2003). Kendall's tau for elliptical distributions. In *Credit Risk* 149–156.

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453

LIU, H., HAN, F. and ZHANG, C.-H. (2012). Transelliptical graphical models. In *Proc. of NIPS* 809–817.

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

LIU, H. and WANG, L. (2017). TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Electron. J. Stat.* **11** 241–294. MR3606771

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

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

MITRA, R. and ZHANG, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Preprint. Available at arXiv:1403.6195.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

SUN, T. and ZHANG, C.-H. (2012a). Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm". *Statist. Sinica* **22** 1354–1358.

SUN, T. and ZHANG, C.-H. (2012b). Sparse matrix inversion with scaled lasso. Preprint. Available at arXiv:1202.2723.

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

WEGKAMP, M. and ZHAO, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612

XUE, L., ZOU, H. and CAI, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.* **40** 1403–1429. MR3015030

YANG, E., ALLEN, G. I., LIU, Z. and RAVIKUMAR, P. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems* 25 1358–1366. Curran Associates, Red Hook, NY.

YANG, E., BAKER, Y., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2014). Mixed graphical models via exponential families. In *Proc. 17th Int. Conf. Artif. Intel. Stat.* 1042–1050.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

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

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHAO, T. and LIU, H. (2014). Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE Trans. Inform. Theory* **60** 7874–7887. MR3285751

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: rina@uchicago.edu

BOOTH SCHOOL OF BUSINESS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: mkolar@chicagobooth.edu