Journal of Machine Learning Research 10 (2009) 2295-2328 Submitted 3/09; Revised 5/09; Published 10/09

The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs

Han Liu HANLIU@CS.CMU.EDU
John Lafferty LAFFERTY@CS.CMU.EDU
Larry Wasserman LARRY@STAT.CMU.EDU
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

Editor: Martin J. Wainwright

Abstract

Recent methods for estimating sparse undirected graphs for real-valued data in high dimensional problems rely heavily on the assumption of normality. We show how to use a semiparametric Gaussian copula—or "nonparanormal"—for high dimensional inference. Just as additive models extend linear models by replacing linear functions with a set of one-dimensional smooth functions, the nonparanormal extends the normal by transforming the variables by smooth functions. We derive a method for estimating the nonparanormal, study the method's theoretical properties, and show that it works well in many examples.

Keywords: graphical models, Gaussian copula, high dimensional inference, sparsity, ℓ_1 regularization, graphical lasso, paranormal, occult

1. Introduction

The linear model is a mainstay of statistical inference that has been extended in several important ways. An extension to high dimensions was achieved by adding a sparsity constraint, leading to the lasso (Tibshirani, 1996). An extension to nonparametric models was achieved by replacing linear functions with smooth functions, leading to additive models (Hastie and Tibshirani, 1999). These two ideas were recently combined, leading to an extension called sparse additive models (SpAM) (Ravikumar et al., 2008, 2009a). In this paper we consider a similar nonparametric extension of undirected graphical models based on multivariate Gaussian distributions in the high dimensional setting. Specifically, we use a high dimensional Gaussian copula with nonparametric marginals, which we refer to as a nonparanormal distribution.

If X is a p-dimensional random vector distributed according to a multivariate Gaussian distribution with covariance matrix Σ, the conditional independence relations between the random variables $X_1, X_2, \ldots, X_p$ are encoded in a graph formed from the precision matrix $\Omega = \Sigma^{-1}$. Specifically, missing edges in the graph correspond to zeroes of Ω. To estimate the graph from a sample of size n, it is only necessary to estimate Σ, which is easy if n is much larger than p. However, when p is larger than n, the problem is more challenging. Recent work has focused on the problem of estimating the graph in this high dimensional setting, which becomes feasible if G is sparse. Yuan and Lin (2007)

Assumptions	Dimension	Regression	Graphical Models
parametric	low	linear model	multivariate normal
	high	lasso	graphical lasso
nonparametric	low	additive model	nonparanormal
	high	sparse additive model	ℓ_1-regularized nonparanormal

Figure 1: Comparison of regression and graphical models. The nonparanormal extends additive models to the graphical model setting. Regularizing the inverse covariance leads to an extension to high dimensions, which parallels sparse additive models for regression.

and Banerjee et al. (2008) propose an estimator based on regularized maximum likelihood using an ℓ_1 constraint on the entries of Ω, and Friedman et al. (2007) develop an efficient algorithm for computing the estimator using a graphical version of the lasso. The resulting estimation procedure has excellent theoretical properties, as shown recently by Rothman et al. (2008) and Ravikumar et al. (2009b).

While Gaussian graphical models can be useful, a reliance on exact normality is limiting. Our goal in this paper is to weaken this assumption. Our approach parallels the ideas behind sparse additive models for regression (Ravikumar et al., 2008, 2009a). Specifically, we replace the Gaussian with a semiparametric Gaussian copula. This means that we replace the random variable $X = (X_1, \ldots, X_p)$ by the transformed random variable $f(X) = (f_1(X_1), \ldots, f_p(X_p))$, and assume that $f(X)$ is multivariate Gaussian. This semiparametric copula results in a nonparametric extension of the normal that we call the *nonparanormal* distribution. The nonparanormal depends on the functions $\{f_j\}$, and a mean μ and covariance matrix Σ, all of which are to be estimated from data. While the resulting family of distributions is much richer than the standard parametric normal (the paranormal), the independence relations among the variables are still encoded in the precision matrix $\Omega = \Sigma^{-1}$. We propose a nonparametric estimator for the functions $\{f_j\}$, and show how the graphical lasso can be used to estimate the graph in the high dimensional setting. The relationship between linear regression models, Gaussian graphical models, and their extensions to nonparametric and high dimensional models is summarized in Figure 1.

Most theoretical results on semiparametric copulas focus on low or at least finite dimensional models (Klaassen and Wellner, 1997; Tsukahara, 2005). Models with increasing dimension require a more delicate analysis; in particular, simply plugging in the usual empirical distribution of the marginals does not lead to accurate inference. Instead we use a truncated empirical distribution. We give a theoretical analysis of this estimator, proving consistency results with respect to risk, model selection, and estimation of Ω in the Frobenius norm.

In the following section we review the basic notion of the graph corresponding to a multivariate Gaussian, and formulate different criteria for evaluating estimators of the covariance or inverse covariance. In Section 3 we present the nonparanormal, and in Section 4 we discuss estimation of the model. We present a theoretical analysis of the estimation method in Section 5, with the detailed proofs collected in an appendix. In Section 6 we present experiments with both simulated data and gene microarray data, where the problem is to construct the isoprenoid biosynthetic pathway.

2. Estimating Undirected Graphs

Let $X = (X_1, \ldots, X_p)$ denote a random vector with distribution $P = N(\mu, \Sigma)$. The undirected graph $G = (V, E)$ corresponding to P consists of a vertex set V and an edge set E. The set V has p elements, one for each component of X. The edge set E consists of ordered pairs (i, j) where $(i, j) \in E$ if there is an edge between X_i and X_j. The edge between (i, j) is excluded from E if and only if X_i is independent of X_j given the other variables $X_{\setminus\{i,j\}} \equiv (X_s : 1 \leq s \leq p, \ s \neq i, j)$, written

$$X_i \perp\!\!\!\perp X_j \,\Big|\, X_{\setminus\{i,j\}}. \tag{1}$$

It is well known that, for multivariate Gaussian distributions, (1) holds if and only if $\Omega_{ij} = 0$ where $\Omega = \Sigma^{-1}$.

Let $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ be a random sample from P, where $X^{(i)} \in \mathbb{R}^p$. If n is much larger than p, then we can estimate Σ using maximum likelihood, leading to the estimate $\widehat{\Omega} = S^{-1}$, where

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(X^{(i)} - \overline{X}\right) \left(X^{(i)} - \overline{X}\right)^T$$

is the sample covariance, with $\overline{X}$ the sample mean. The zeroes of Ω can then be estimated by applying hypothesis testing to $\widehat{\Omega}$ (Drton and Perlman, 2007, 2008).

When $p > n$, maximum likelihood is no longer useful; in particular, the estimate $\widehat{\Sigma}$ is not positive definite, having rank no greater than n. Inspired by the success of the lasso for linear models, several authors have suggested estimating Σ by minimizing

$$-\ell(\Omega) + \lambda \sum_{j \neq k} |\Omega_{jk}|$$

where

$$\ell(\Omega) = \frac{1}{2} \left(\log |\Omega| - \mathrm{tr}(\Omega S) - p \log(2\pi)\right)$$

is the log-likelihood with S the sample covariance matrix. The estimator $\widehat{\Omega}$ can be computed efficiently using the glasso algorithm (Friedman et al., 2007), which is a block coordinate descent algorithm that uses the standard lasso to estimate a single row and column of Ω in each iteration. Under appropriate sparsity conditions, the resulting estimator $\widehat{\Omega}$ has been shown to have good theoretical properties (Rothman et al., 2008; Ravikumar et al., 2009b).

There are several different ways to judge the quality of an estimator $\widehat{\Sigma}$ of the covariance or $\widehat{\Omega}$ of the inverse covariance. We discuss three in this paper, persistency, norm consistency, and sparsistency. Persistency means consistency in risk, when the model is not necessarily assumed to be correct. Suppose the true distribution P has mean μ_0, and that we use a multivariate normal $p(x; \mu_0, \Sigma)$ for prediction; we do not assume that P is normal. We observe a new vector $X \sim P$ and define the prediction risk to be

$$R(\Sigma) = -\mathbb{E} \log p(X; \mu_0, \Sigma) = -\int \log p(x; \mu_0, \Sigma)\, dP(x).$$

It follows that

$$R(\Sigma) = \frac{1}{2} \left(\mathrm{tr}(\Sigma^{-1}\Sigma_0) + \log |\Sigma| - p \log(2\pi)\right)$$

where Σ_0 is the covariance of X under P. If $\mathcal{S}$ is a set of covariance matrices, the oracle is defined to be the covariance matrix Σ_* that minimizes $R(\Sigma)$ over $\mathcal{S}$:

$$\Sigma_* = \arg\min_{\Sigma \in \mathcal{S}} R(\Sigma).$$

Thus $p(x;\mu_0,\Sigma_*)$ is the best predictor of a new observation among all distributions in $\{p(x;\mu_0,\Sigma) : \Sigma \in \mathcal{S}\}$. In particular, if $\mathcal{S}$ consists of covariance matrices with sparse graphs, then $p(x;\mu_0,\Sigma_*)$ is, in some sense, the best sparse predictor. An estimator $\widehat{\Sigma}_n$ is *persistent* if

$$R(\widehat{\Sigma}_n) - R(\Sigma_*) \xrightarrow{P} 0$$

as the sample size n increases to infinity. Thus, a persistent estimator approximates the best estimator over the class $\mathcal{S}$, but we do not assume that the true distribution has a covariance matrix in $\mathcal{S}$, or even that it is Gaussian. Moreover, we allow the dimension $p = p_n$ to increase with n. On the other hand, norm consistency and sparsistency require that the true distribution is Gaussian. In this case, let Σ_0 denote the true covariance matrix. An estimator is *norm consistent* if

$$\|\widehat{\Sigma}_n - \Sigma\| \xrightarrow{P} 0$$

where $\|\cdot\|$ is a norm. If $E(\Omega)$ denotes the edge set corresponding to Ω, an estimator is *sparsistent* if

$$\mathbb{P}\Big(E(\Omega) \neq E(\widehat{\Omega}_n)\Big) \to 0.$$

Thus, a sparsistent estimator identifies the correct graph consistently. We present our theoretical analysis on these properties of the nonparanormal in Section 5.

3. The Nonparanormal

We say that a random vector $X = (X_1,\ldots,X_p)^T$ has a *nonparanormal* distribution if there exist functions $\{f_j\}_{j=1}^p$ such that $Z \equiv f(X) \sim N(\mu,\Sigma)$, where $f(X) = (f_1(X_1),\ldots,f_p(X_p))$. We then write

$$X \sim NPN(\mu,\Sigma,f).$$

When the f_j's are monotone and differentiable, the joint probability density function of X is given by

$$p_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(f(x)-\mu)^T \Sigma^{-1}(f(x)-\mu)\right\} \prod_{j=1}^p |f_j'(x_j)|. \tag{2}$$

Lemma 1 *The nonparanormal distribution $NPN(\mu,\Sigma,f)$ is a Gaussian copula when the f_j's are monotone and differentiable.*

Proof By Sklar's theorem (Sklar, 1959), any joint distribution can be written as

$$F(x_1,\ldots,x_p) = C\{F_1(x_1),\ldots,F_p(x_p)\}$$

where the function C is called a copula. For the nonparanormal we have

$$F(x_1,\ldots,x_p) = \Phi_{\mu,\Sigma}(\Phi^{-1}(F_1(x_1)),\ldots,\Phi^{-1}(F_p(x_p)))$$

where $\Phi_{\mu,\Sigma}$ is the multivariate Gaussian cdf and Φ is the univariate standard Gaussian cdf. Thus, the corresponding copula is

$$C(u_1,\ldots,u_p) = \Phi_{\mu,\Sigma}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_p)).$$

This is exactly a Gaussian copula with parameters μ and Σ. If each f_j is differentiable then the density of X has the same form as (2). ■

Note that the density in (2) is not identifiable; to make the family identifiable we demand that f_j preserve means and variances:

$$\mu_j = \mathbb{E}(Z_j) = \mathbb{E}(X_j) \text{ and } \sigma_j^2 \equiv \Sigma_{jj} = \mathrm{Var}(Z_j) = \mathrm{Var}(X_j). \tag{3}$$

Note that these conditions only depend on $\mathrm{diag}(\Sigma)$ but not the full covariance matrix.

Let $F_j(x)$ denote the marginal distribution function of X_j. Then

$$F_j(x) = \mathbb{P}(X_j \le x) = \mathbb{P}(Z_j \le f_j(x)) = \Phi\left(\frac{f_j(x)-\mu_j}{\sigma_j}\right)$$

which implies that

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x)). \tag{4}$$

The following basic fact says that the independence graph of the nonparanormal is encoded in $\Omega = \Sigma^{-1}$, as for the parametric normal.

Lemma 2 *If $X \sim NPN(\mu,\Sigma,f)$ is nonparanormal and each f_j is differentiable, then $X_i \perp\!\!\!\perp X_j \,|\, X_{\setminus\{i,j\}}$ if and only if $\Omega_{ij} = 0$, where $\Omega = \Sigma^{-1}$.*

Proof From the form of the density (2), it follows that the density factors with respect to the graph of Ω, and therefore obeys the global Markov property of the graph. ■

Next we show that the above is true for any choice of identification restrictions.

Lemma 3 *Define*

$$h_j(x) = \Phi^{-1}(F_j(x)) \tag{5}$$

and let Λ be the covariance matrix of $h(X)$. Then $X_j \perp\!\!\!\perp X_k \,|\, X_{\setminus\{j,k\}}$ if and only if $\Lambda^{-1}_{jk} = 0$.

Proof We can rewrite the covariance matrix as

$$\Sigma_{jk} = \mathrm{Cov}(Z_j,Z_k) = \sigma_j\sigma_k \mathrm{Cov}(h_j(X_j),h_k(X_k)).$$

Hence $\Sigma = D\Lambda D$ and

$$\Sigma^{-1} = D^{-1}\Lambda^{-1}D^{-1},$$

where D is the diagonal matrix with $\mathrm{diag}(D) = \sigma$. The zero pattern of Λ^{-1} is therefore identical to the zero pattern of Σ^{-1}. ■

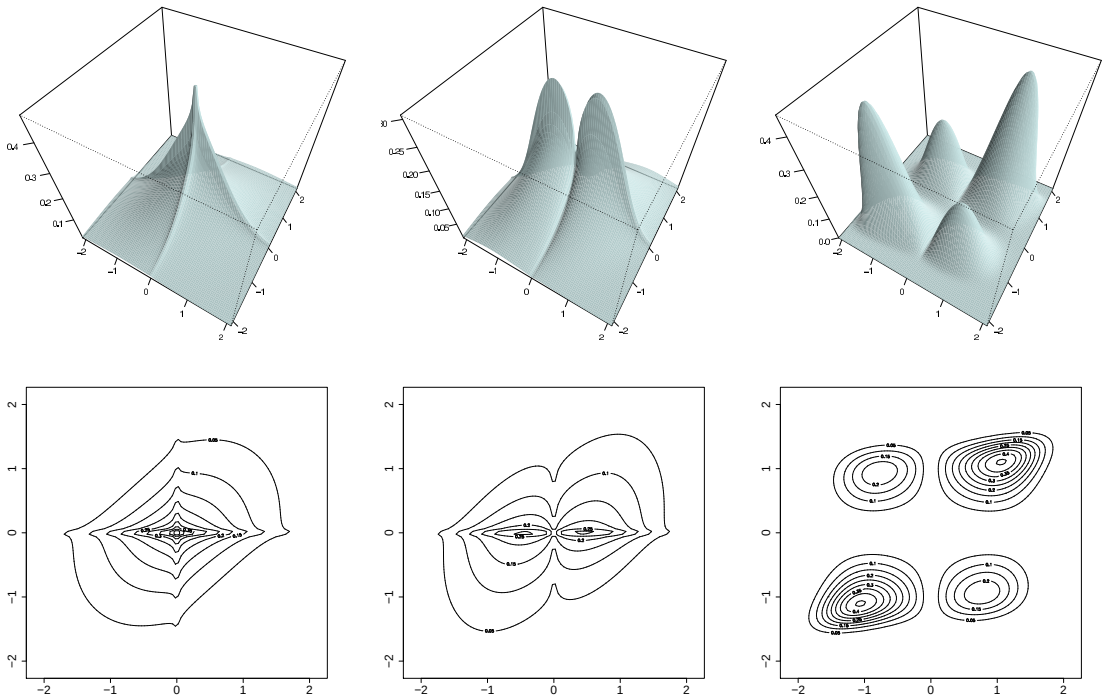


Figure 2: Densities of three 2-dimensional nonparanormals. The component functions have the form $f_j(x) = \text{sign}(x)|x|^{\alpha_j}$. Left: $\alpha_1 = 0.9$, $\alpha_2 = 0.8$; center: $\alpha_1 = 1.2$, $\alpha_2 = 0.8$; right $\alpha_1 = 2$, $\alpha_2 = 3$. In each case $\mu = (0,0)$ and $\Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$.

Thus, it is not necessary to estimate μ or σ to estimate the graph.

Figure 2 shows three examples of 2-dimensional nonparanormal densities. In each case, the component functions $f_j(x)$ take the form

$$f_j(x) = a_j \text{sign}(x)|x|^{\alpha_j} + b_j$$

where the constants a_j and b_j are set to enforce the identifiability constraints (3). The covariance in each case is $\Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$ and the mean is $\mu = (0,0)$. The exponent α_j determines the nonlinearity. It can be seen how the concavity of the density changes with the exponent α, and that $\alpha > 1$ can result in multiple modes.

The assumption that $f(X) = (f_1(X_1), \ldots, f_p(X_p)$ is normal leads to a semiparametric model where only one dimensional functions need to be estimated. But the monotonicity of the functions f_j, which map onto $\mathbb{R}$, enables computational tractability of the nonparanormal. For more general functions f, the normalizing constant for the density

$$p_X(x) \propto \exp\left\{ -\frac{1}{2} (f(x) - \mu)^T \Sigma^{-1} (f(x) - \mu) \right\}$$

cannot be computed in closed form.

4. Estimation Method

Let $X^{(1)},\ldots,X^{(n)}$ be a sample of size n where $X^{(i)} = (X_1^{(i)},\ldots,X_p^{(i)})^T \in \mathbb{R}^p$. In light of (5) we define

$$\widehat{h}_j(x) = \Phi^{-1}(\widetilde{F}_j(x))$$

where $\widetilde{F}_j$ is an estimator of F_j. A natural candidate for $\widetilde{F}_j$ is the marginal empirical distribution function

$$\widehat{F}_j(t) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{X_j^{(i)} \le t\right\}}.$$

Now, let θ denote the parameters of the copula. Tsukahara (2005) suggests taking $\widehat{\theta}$ to be the solution of

$$\sum_{i=1}^{n} \phi\left(\widetilde{F}_1(X_1^{(i)}),\ldots,\widetilde{F}_p(X_p^{(i)}),\theta\right) = 0$$

where ϕ is an estimating equation and $\widetilde{F}_j(t) = n\widehat{F}_j(t)/(n+1)$. In our case, θ corresponds to the covariance matrix. The resulting estimator $\widehat{\theta}$, called a rank approximate Z-estimator, has excellent theoretical properties. However, we are interested in the high dimensional scenario where the dimension p is allowed to increase with n; the variance of $\widehat{F}_j(t)$ is too large in this case. Instead, we use the following truncated or *Winsorized*[1] estimator:

$$\widetilde{F}_j(x) = \begin{cases} \delta_n & \text{if } \widehat{F}_j(x) < \delta_n \\ \widehat{F}_j(x) & \text{if } \delta_n \le \widehat{F}_j(x) \le 1-\delta_n \\ (1-\delta_n) & \text{if } \widehat{F}_j(x) > 1-\delta_n, \end{cases} \tag{6}$$

where δ_n is a truncation parameter. Clearly, there is a bias-variance tradeoff in choosing δ_n. Essentially the same estimator with $\delta_n = 1/n$ is studied by Klaassen and Wellner (1997) in the case of bivariate Gaussian copula. In what follows we use

$$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}.$$

This provides the right balance so that we can achieve the desired rate of convergence in our estimate of Ω and the associated undirected graph G in the high dimensional setting.

Given this estimate of the distribution of variable X_j, we then estimate the transformation function f_j by

$$\widetilde{f}_j(x) \equiv \widehat{\mu}_j + \widehat{\sigma}_j \widetilde{h}_j(x) \tag{7}$$

where

$$\widetilde{h}_j(x) = \Phi^{-1}\left(\widetilde{F}_j(x)\right)$$

and $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ are the sample mean and the standard deviation:

$$\widehat{\mu}_j \equiv \frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} \text{ and } \widehat{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_j^{(i)} - \widehat{\mu}_j\right)^2}.$$

1. After Charles P. Winsor, whom John Tukey credited with converting him from topology to statistics Mallows 1990.

Now, let $S_n(\widetilde{f})$ be the sample covariance matrix of $\widetilde{f}(X^{(1)}),\ldots,\widetilde{f}(X^{(n)})$; that is,

$$\begin{aligned} S_n(\widetilde{f}) &\equiv \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{f}(X^{(i)})-\mu_n(\widetilde{f})\right)\left(\widetilde{f}(X^{(i)})-\mu_n(\widetilde{f})\right)^T \qquad (8)\\ \mu_n(\widetilde{f}) &\equiv \frac{1}{n}\sum_{i=1}^{n}\widetilde{f}(X^{(i)}). \end{aligned}$$

We then estimate Ω using $S_n(\widetilde{f})$. For instance, the maximum likelihood estimator is $\widehat{\Omega}_n^{\mathrm{MLE}} = S_n(\widetilde{f})^{-1}$. The ℓ_1-regularized estimator is

$$\widehat{\Omega}_n = \arg\min_{\Omega}\left\{\mathrm{tr}\left(\Omega S_n(\widetilde{f})\right)-\log|\Omega|+\lambda\|\Omega\|_1\right\} \qquad (9)$$

where λ is a regularization parameter, and $\|\Omega\|_1=\sum_{j\neq k}|\Omega_{jk}|$. The estimated graph is then $\widehat{E}_n = \{(j,k):\ \widehat{\Omega}_{jk}\neq 0\}$.

The nonparanormal is analogous to a sparse additive regression model (Ravikumar et al., 2009a), in the sense that both methods transform the variables by univariate functions. However, while sparse additive models use a regularized risk criterion to fit univariate transformations, our nonparanormal estimator uses a two-step procedure:

1. Replace the observations, for each variable, by their respective normal scores, subject to a Winsorized truncation.

2. Apply the graphical lasso to the transformed data to estimate the undirected graph.

The first step is non-iterative and computationally efficient, with no tuning parameters; it also makes the nonparanormal amenable to theoretical analysis.

Starting with the model in (2), another possibility would be to parametrize each f_j according to some parametric class of monotone functions such as the Box-Cox family, and then find the maximum likelihood estimates of $(\Omega, f_1,\ldots f_p)$ in that class. This might lead to estimates of f_j that depend on Ω, and vice versa, and the estimation problem would not in general be convex. Alternatively, due to (4), the marginal information could be used to estimate the parameters. Our nonparametric approach to estimating the transformations has the advantages of making few assumptions and being easy to compute. In the following section we analyze the theoretical properties of this estimator.

5. Theoretical Results

In this section we present our theoretical results on risk consistency, model selection consistency, and norm consistency of the covariance Σ and inverse covariance Ω. From Lemma 3, the estimate of the graph does not depend on $\sigma_j,\ j\in\{1,\ldots,p\}$ and μ, so we assume that $\sigma_j=1$ and $\mu=0$. Our key technical result is an analysis of the covariance of the Winsorized estimator defined in (6), (7), and (8). In particular, we show that under appropriate conditions,

$$\max_{j,k}\left|S_n(\widetilde{f})_{jk}-S_n(f)_{jk}\right| = o_P(1)$$

where $S_n(\widetilde{f})_{jk}$ denotes the (j,k) entry of the matrix. This result allows us to leverage the recent analysis of Rothman et al. (2008) and Ravikumar et al. (2009b) in the Gaussian case to obtain consistency results for the nonparanormal. More precisely, our main theorem is the following.

Theorem 4 *Suppose that $p = n^{\xi}$ and let $\widetilde{f}$ be the Winsorized estimator defined in* (7) *with* $\delta_n = \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$. *Define*

$$C_M \equiv \frac{48}{\sqrt{\pi}}\left(\sqrt{2M}-1\right)(M+2). \tag{10}$$

For some $M \geq 2(\xi+1)$.

Then for any $\varepsilon \geq C_M \sqrt{\frac{\log p \log^2 n}{n^{1/2}}}$ *and sufficiently large n, we have*

$$\begin{aligned}
&\mathbb{P}\left(\max_{jk}\left|S_n(\widetilde{f})_{jk} - S_n(f)_{jk}\right| > 2\varepsilon\right) \\
&\quad\leq \quad \frac{1}{2\sqrt{\pi \log(np)}} + 2\exp\left(2\log p - \frac{n^{1/2}\varepsilon^2}{1232\pi^2 \log^2 n}\right) + 2\exp\left(2\log p - \frac{n^{1/2}}{8\pi \log n}\right) + o(1).
\end{aligned}$$

The proof of the above theorem is given in Section 7. The following corollary is immediate, and specifies the scaling of the dimension in terms of sample size.

Corollary 5 *Let* $M \geq \max\{15\pi, 2\xi+1\}$. *Then*

$$\mathbb{P}\left(\max_{jk}\left|S_n(\widetilde{f})_{jk} - S_n(f)_{jk}\right| > 2C_M\sqrt{\frac{\log p \log^2 n}{n^{1/2}}}\right) = o(1).$$

Hence,

$$\max_{j,k}\left|S_n(\widetilde{f})_{jk} - S_n(f)_{jk}\right| = O_P\left(\sqrt{\frac{\log p \log^2 n}{n^{1/2}}}\right).$$

The following corollary yields estimation consistency in both the Frobenius norm and the ℓ_2-operator norm. The proof follows the same arguments as the proof of Theorem 1 and Theorem 2 from Rothman et al. (2008), replacing their Lemma 1 with our Theorem 4.

For a matrix $A = (a_{ij})$, the Frobenius norm $\|\cdot\|_F$ is defined as $\|A\|_{\mathrm{F}} \equiv \sqrt{\sum_{i,j} a_{ij}^2}$. The ℓ_2-operator norm $\|\cdot\|_2$ is defined as the magnitude of the largest eigenvalue of the matrix, $\|A\|_2 \equiv \max_{\|x\|_2=1}\|Ax\|_2$. In the following, we write $a_n \asymp b_n$ if there are positive constants c and C independent of n such that $c \leq a_n/b_n \leq C$.

Corollary 6 *Suppose that the data are generated as* $X^{(i)} \sim NPN(\mu_0, \Sigma_0, f_0)$, *and let* $\Omega_0 = \Sigma_0^{-1}$. *If the regularization parameter* λ_n *is chosen as*

$$\lambda_n \asymp 2C_M\sqrt{\frac{\log p \log^2 n}{n^{1/2}}}$$

where C_M *is defined in Theorem 4. Then the nonparanormal estimator* $\widehat{\Omega}_n$ *of* (9) *satisfies*

$$\|\widehat{\Omega}_n - \Omega_0\|_{\mathrm{F}} = O_P\left(\sqrt{\frac{(s+p)(\log p \log^2 n)}{n^{1/2}}}\right)$$

and

$$\|\widehat{\Omega}_n - \Omega_0\|_2 = O_P\left(\sqrt{\frac{s(\log p \log^2 n)}{n^{1/2}}}\right),$$

where

$$s \equiv \text{Card}\left(\{(i,j) \in \{1,\ldots,p\} \times \{1,\ldots,p\} \,|\, \Omega_0(i,j) \neq 0,\ i \neq j\}\right)$$

is the number of nonzero off-diagonal elements of the true precision matrix.

To prove the model selection consistency result, we need further assumptions. We follow Ravikumar (2009) and let the $p^2 \times p^2$ Fisher information matrix of Σ_0 be $\Gamma \equiv \Sigma_0 \otimes \Sigma_0$ where $\otimes$ is the Kronecker matrix product, and define the support set S of $\Omega_0 = \Sigma_0^{-1}$ as

$$S \equiv \{(i,j) \in \{1,\ldots,p\} \times \{1,\ldots,p\} \,|\, \Omega_0(i,j) \neq 0\}.$$

We use S^c to denote the complement of S in the set $\{1,\ldots,p\} \times \{1,\ldots,p\}$, and for any two subsets T and T' of $\{1,\ldots,p\} \times \{1,\ldots,p\}$, we use $\Gamma_{TT'}$ to denote the sub-matrix with rows and columns of Γ indexed by T and T' respectively.

Assumption 1 *There exists some $\alpha \in (0,1]$, such that $\left\|\Gamma_{S^cS}(\Gamma_{SS})^{-1}\right\|_\infty \leq 1-\alpha$.*

As in Ravikumar et al. (2009b), we define two quantities $K_{\Sigma_0} \equiv \|\Sigma_0\|_\infty$ and $K_\Gamma \equiv \|(\Gamma_{SS})^{-1}\|_\infty$. Further, we define the maximum row degree as

$$d \equiv \max_{i=1,\ldots,p} \text{Card}\left(\{j \in 1,\ldots,p \,|\, \Omega_0(i,j) \neq 0\}\right).$$

Assumption 2 *The quantities K_{Σ^0} and K_Γ are bounded, and there are positive constants C such that*

$$\min_{(j,k)\in S} |\Omega_0(j,k)| \geq C\sqrt{\frac{\log^3 n}{n^{1/2}}}$$

for large enough n.

The proof of the following corollary uses our Theorem 4 in place of Equation (12) in the analysis of Ravikumar et al. (2009b).

Corollary 7 *Suppose the regularization parameter is chosen as*

$$\lambda_n \asymp 2C_M\sqrt{\frac{\log p \log^2 n}{n^{1/2}}}$$

where $C(M,n,p)$ is defined in Theorem 4. Then the nonparanormal estimator $\widehat{\Omega}_n$ satisfies

$$\mathbb{P}\left(\mathcal{G}\left(\widehat{\Omega}_n, \Omega_0\right)\right) \geq 1 - o(1)$$

where $\mathcal{G}(\widehat{\Omega}_n, \Omega_0)$ is the event

$$\left\{\text{sign}\left(\widehat{\Omega}_n(j,k)\right) = \text{sign}\left(\Omega_0(j,k)\right), \quad \forall j,k \in S\right\}.$$

Our persistency (risk consistency) result parallels the persistency result for additive models given in Ravikumar et al. (2009a), and allows model dimension that grows exponentially with sample size. The definition in this theorem uses the fact (from Lemma 11) that $\sup_x \Phi^{-1}\left(\widetilde{F}_j(x)\right) \leq \sqrt{2\log n}$ when $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$.

In the next theorem, we do not assume the true model is nonparanormal and define the population and sample risks as

$$\begin{aligned} R(f,\Omega) &= \frac{1}{2}\left\{\mathrm{tr}\left[\Omega \mathbb{E}(f(X)f(X)^T\right] - \log|\Omega| - p\log(2\pi)\right\} \\ \widehat{R}(f,\Omega) &= \frac{1}{2}\left\{\mathrm{tr}\left[\Omega S_n(f)\right] - \log|\Omega| - p\log(2\pi)\right\}. \end{aligned}$$

Theorem 8 *Suppose that $p \leq e^{n^\xi}$ for some $\xi < 1$, and define the classes*

$$\begin{aligned} \mathcal{M}_n &= \left\{f : \mathbb{R} \to \mathbb{R} : f \text{ is monotone with } \|f\|_\infty \leq C\sqrt{\log n}\right\} \\ \mathcal{C}_n &= \left\{\Omega : \|\Omega^{-1}\|_1 \leq L_n\right\}. \end{aligned}$$

Let $\widehat{\Omega}_n$ be given by

$$\widehat{\Omega}_n = \underset{\Omega \in \mathcal{C}_n}{\arg\min}\left\{tr\left(\Omega S_n(\widetilde{f})\right) - \log|\Omega|\right\}.$$

Then

$$R(\widetilde{f}_n, \widehat{\Omega}_n) - \inf_{(f,\Omega)\in \mathcal{M}_n^p \oplus \mathcal{C}_n} R(f,\Omega) = O_P\left(L_n\sqrt{\frac{\log n}{n^{1-\xi}}}\right).$$

Hence the Winsorized estimator of (f,Ω) with $\delta_n = 1/(4n^{1/4}\sqrt{\pi\log n})$ is persistent over $\mathcal{C}_n$ when $L_n = o\left(n^{(1-\xi)/2}/\sqrt{\log n}\right)$.

The proofs of Theorems 4 and 8 are given in Section 7.

6. Experimental Results

In this section, we report experimental results on synthetic and real data sets. We mainly compare the ℓ_1-regularized nonparanormal and Gaussian (paranormal) models, computed using the graphical lasso algorithm (glasso) of Friedman et al. (2007). The primary conclusions are: (i) When the data are multivariate Gaussian, the performance of the two methods is comparable; (ii) when the model is correct, the nonparanormal performs much better than the graphical lasso in many cases; (iii) for a particular gene microarray data set, our method behaves differently from the graphical lasso, and may support different biological conclusions.

Note that we can reuse the glasso implementation to fit a sparse nonparanormal. In particular, after computing the Winsorized sample covariance $S_n(\widetilde{f})$, we pass this matrix to the glasso routine to carry out the optimization

$$\widehat{\Omega}_n = \underset{\Omega}{\arg\min}\left\{\mathrm{tr}\left(\Omega S_n(\widetilde{f})\right) - \log|\Omega| + \lambda_n\|\Omega\|_1\right\}.$$

6.1 Neighborhood Graphs

We begin by describing a procedure to generate graphs as in (Meinshausen and Bühlmann, 2006), with respect to which several distributions can then be defined. We generate a p-dimensional sparse graph $G \equiv (V,E)$ as follows: Let $V = \{1,\ldots,p\}$ correspond to variables $X = (X_1,\ldots,X_p)$. We associate each index j with a point $(Y_j^{(1)},Y_j^{(2)}) \in [0,1]^2$ where

$$Y_1^{(k)},\ldots,Y_n^{(k)} \sim \text{Uniform}[0,1]$$

for $k = 1,2$. Each pair of nodes (i,j) is included in the edge set E with probability

$$\mathbb{P}\Big((i,j) \in E\Big) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|y_i - y_j\|_n^2}{2s}\right)$$

where $y_i \equiv (y_i^{(1)},y_i^{(2)})$ is the observation of $(Y_i^{(1)},Y_i^{(2)})$ and $\|\cdot\|_n$ represents the Euclidean distance. Here, $s = 0.125$ is a parameter that controls the sparsity level of the generated graph. We restrict the maximum degree of the graph to be four and build the inverse covariance matrix Ω_0 according to

$$\Omega_0(i,j) = \begin{cases} 1 & \text{if } i = j \\ 0.245 & \text{if } (i,j) \in E \\ 0 & \text{otherwise,} \end{cases}$$

where the value 0.245 guarantees positive definiteness of the inverse covariance matrix.

Given Ω_0, n data points are sampled from

$$X^{(1)},\ldots,X^{(n)} \sim NPN(\mu_0,\Sigma_0,f_0)$$

where $\mu_0 = (1.5,\ldots,1.5)$, $\Sigma_0 = \Omega_0^{-1}$. For simplicity, the transformation functions for all dimensions are the same, $f_1 = \ldots = f_p = f$. To sample data from the nonparanormal distribution, we also require $g \equiv f^{-1}$; two different transformations g are employed.

Definition 9 (Gaussian CDF Transformation) *Let g_0 be a one-dimensional Gaussian cumulative distribution function with mean μ_{g_0} and the standard deviation σ_{g_0}, that is,*

$$g_0(t) \equiv \Phi\left(\frac{t-\mu_{g_0}}{\sigma_{g_0}}\right).$$

We define the transformation function $g_j = f_j^{-1}$ for the j-th dimension as

$$g_j(z_j) \equiv \sigma_j \left(\frac{g_0(z_j) - \int g_0(t)\phi\left(\frac{t-\mu_j}{\sigma_j}\right)dt}{\sqrt{\int \left(g_0(y) - \int g_0(t)\phi\left(\frac{t-\mu_j}{\sigma_j}\right)dt\right)^2 \phi\left(\frac{y-\mu_j}{\sigma_j}\right)dy}} \right) + \mu_j$$

where $\sigma_j = \Sigma_0(j,j)$.

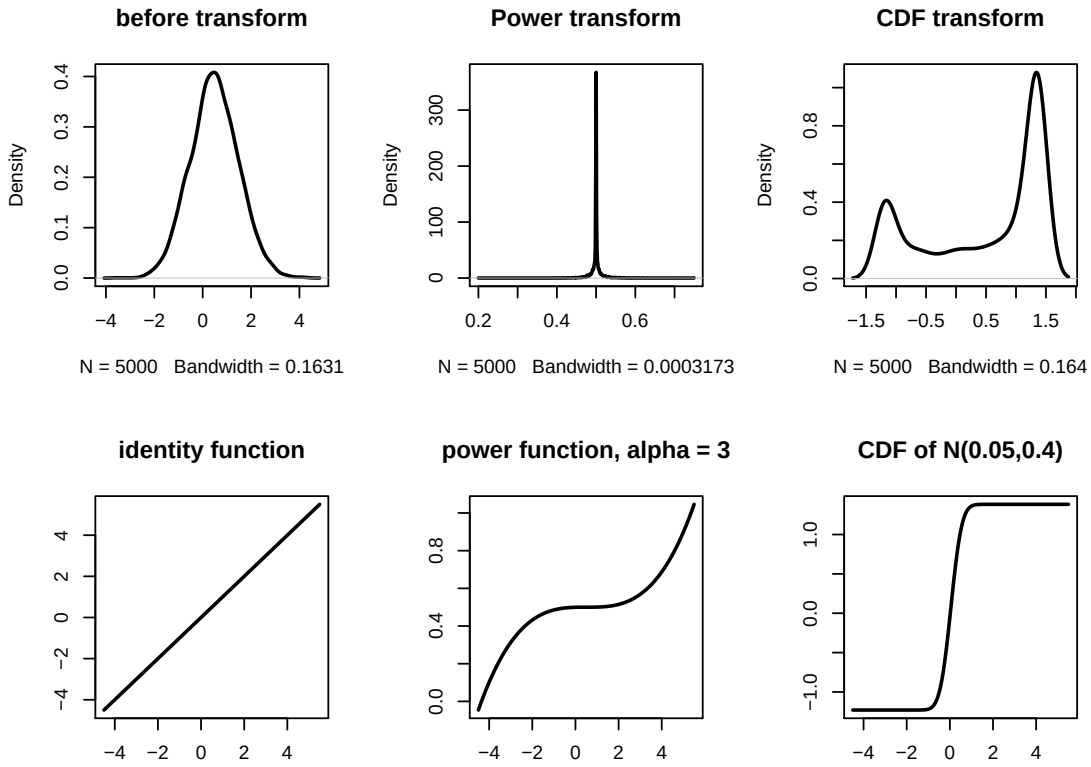


Figure 3: The power and cdf transformations. The densities are estimated using a kernel density estimator with bandwidths selected by cross-validation.

Definition 10 (Symmetric Power Transformation) *Let g_0 be the symmetric and odd transformation given by*

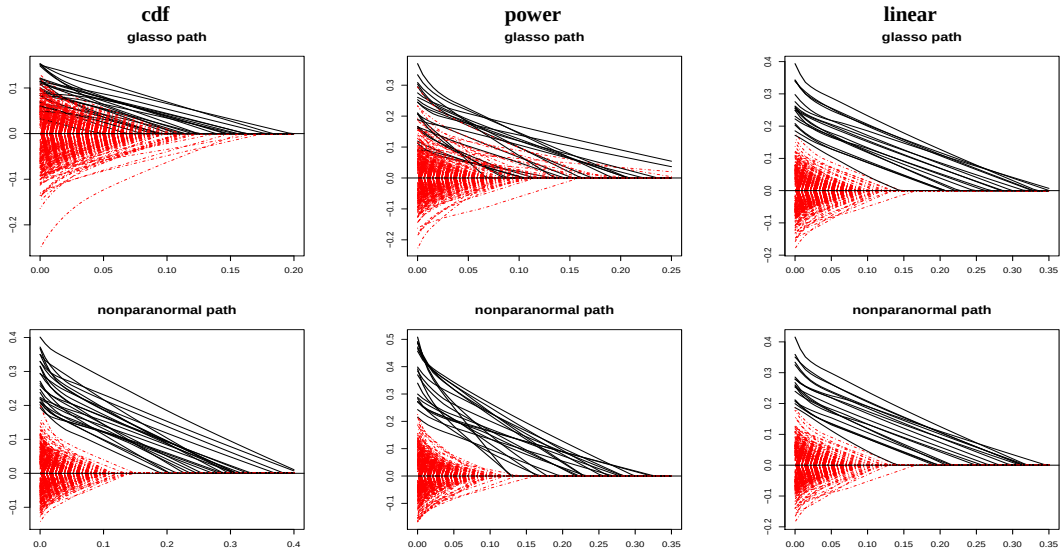
$$g_0(t) = \text{sign}(t)|t|^\alpha$$

where $\alpha > 0$ is a parameter. We define the power transformation for the j-th dimension as

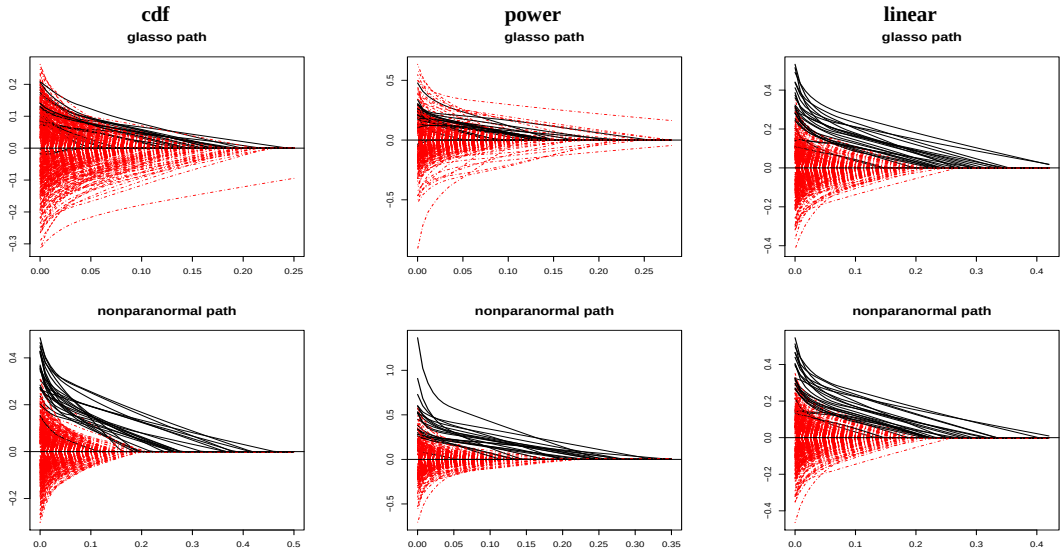
$$g_j(z_j) \equiv \sigma_j \left(\frac{g_0(z_j - \mu_j)}{\sqrt{\int g_0^2(t-\mu_j)\phi\left(\frac{t-\mu_j}{\sigma_j}\right) dt}} \right) + \mu_j.$$

These transformation are constructed to preserve the marginal mean and standard deviation. In the following experiments, we refer to them as the cdf transformation and the power transformation, respectively. For the cdf transformation, we set $\mu_{g_0} = 0.05$ and $\sigma_{g_0} = 0.4$. For the power transformation, we set $\alpha = 3$.

To visualize these two transformations, we sample 5000 data points from a one-dimensional normal distribution $N(0.5, 1.0)$ and then apply the above two transformations; the results are shown in Figure 3. It can be seen how the cdf and power transformations map a univariate normal distribution into a highly skewed and a bi-modal distribution, respectively.



$n = 500$



$n = 200$

Figure 4: Regularization paths for the glasso and nonparanormal with $n = 500$ (top) and $n = 200$ (bottom). The paths for the relevant variables (nonzero inverse covariance entries) are plotted as solid (black) lines; the paths for the irrelevant variables are plotted as dashed (red) lines. For non-Gaussian distributions, the nonparanormal better separates the relevant and irrelevant dimensions.

To generate synthetic data, we set $p = 40$, resulting in $\binom{40}{2} + 40 = 820$ parameters to be estimated, and vary the sample sizes from $n = 200$ to $n = 1000$. Three conditions are considered, corresponding to using the cdf transform, the power transform, or no transformation. In each case, both the glasso and the nonparanormal are applied to estimate the graph.

6.1.1 COMPARISON OF REGULARIZATION PATHS

We choose a set of regularization parameters Λ; for each $\lambda \in \Lambda$, we obtain an estimate $\widehat{\Omega}_n$ which is a 40×40 matrix. The upper triangular matrix has 780 parameters; we vectorize it to get a 780-dimensional parameter vector. A regularization path is the trace of these parameters over all the regularization parameters within Λ. The regularization paths for both methods are plotted in Figure 4. For the cdf transformation and the power transformation, the nonparanormal separates the relevant and the irrelevant dimensions very well. For the glasso, relevant variables are mixed with irrelevant variables. If no transformation is applied, the paths for both methods are almost the same.

6.1.2 ESTIMATED TRANSFORMATIONS

For sample size $n = 1000$, we plot the estimated transformations for three of the variables in Figure 5. It is clear that Winsorization plays a significant role for the power transformation. This is intuitive due to the high skewness of the nonparanormal distribution in this case.

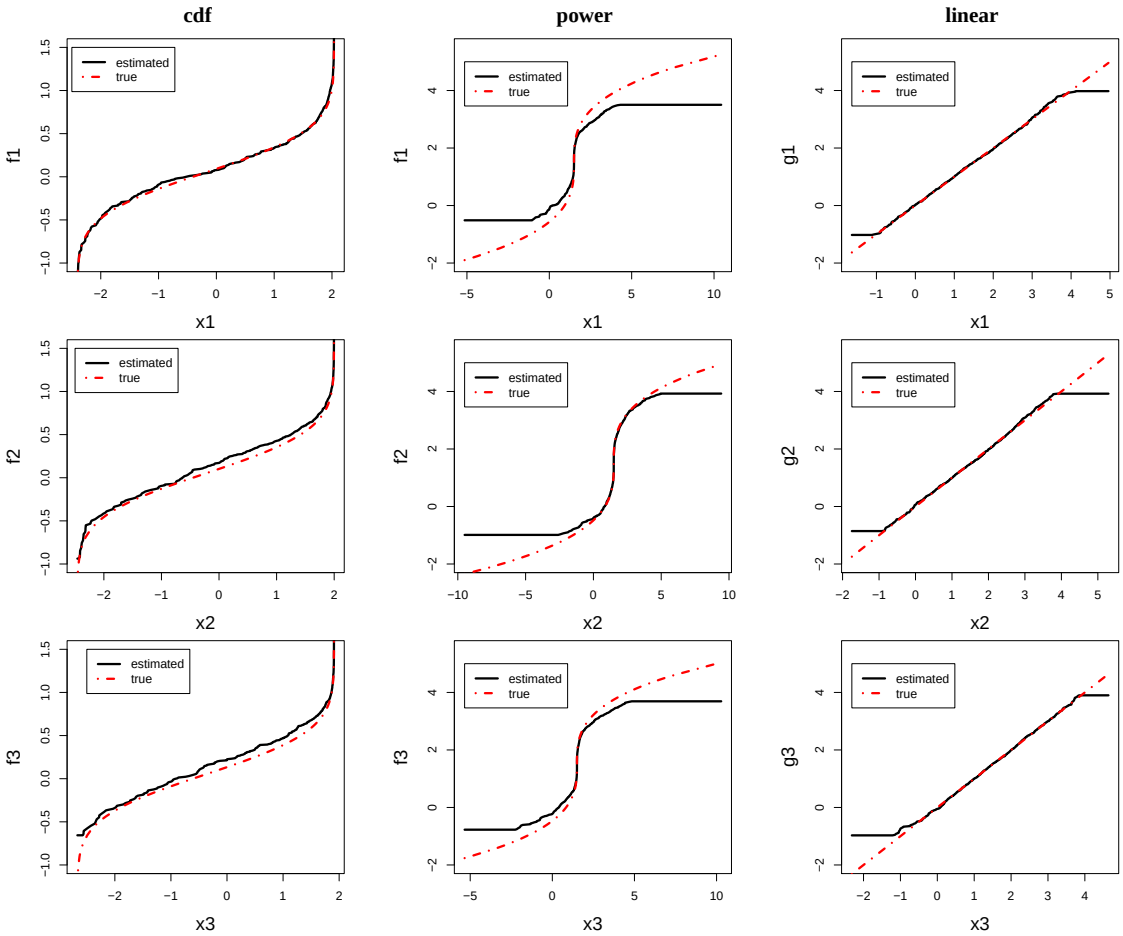


Figure 5: Estimated transformations for the first three variables. Winsorization plays a significant role for the power transformation due to its high skewness.

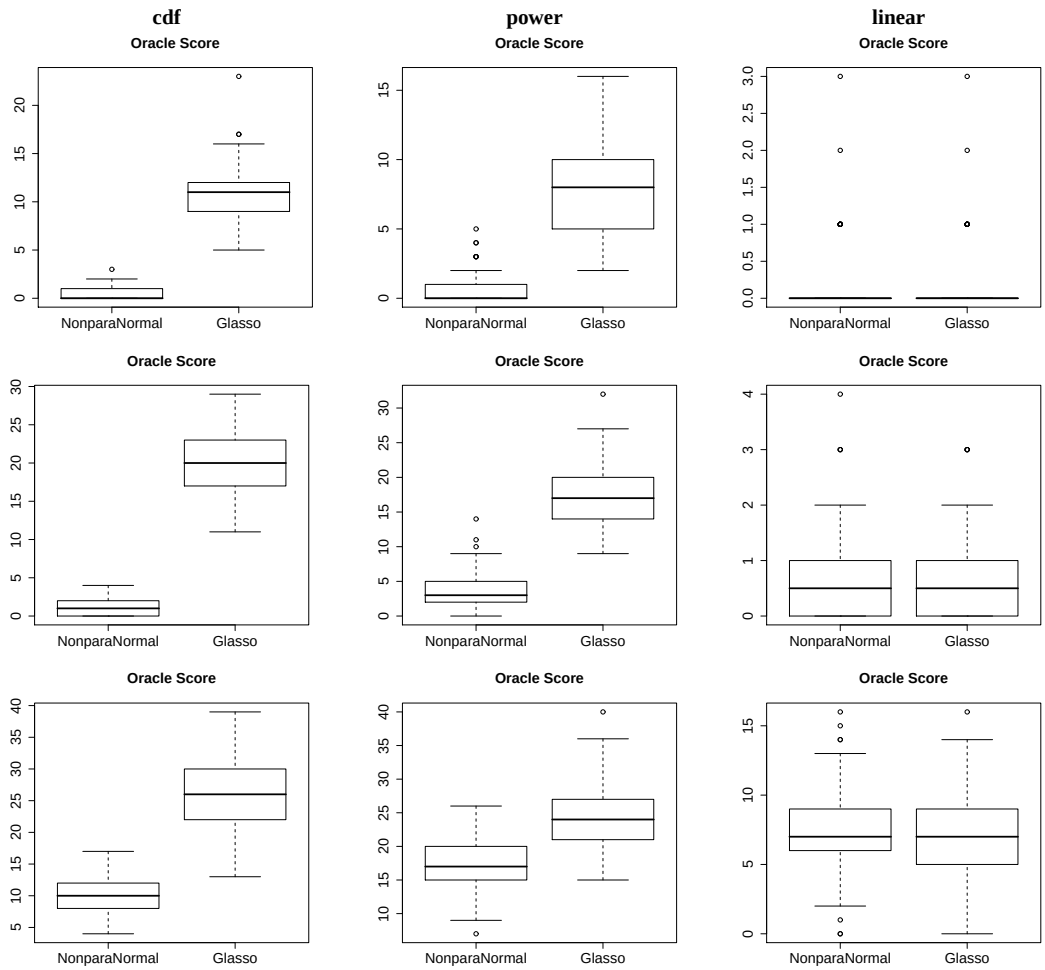


Figure 6: Boxplots of the oracle scores for $n = 1000, 500, 200$ (top, center, bottom).

6.1.3 QUANTITATIVE COMPARISON

To evaluate the performance for structure estimation quantitatively, we use false positive and false negative rates. Let $G = (V, E)$ be a p-dimensional graph (which has at most $\binom{p}{2}$ edges) in which there are $|E| = r$ edges, and let $\widehat{G}^{\lambda} = (V, \widehat{E}^{\lambda})$ be an estimated graph using the regularization parameter λ. The number of false positives at λ is

$$\mathrm{FP}(\lambda) \equiv \text{number of edges in } \widehat{E}^{\lambda} \text{ not in } E$$

The number of false negatives at λ is defined as

$$\mathrm{FN}(\lambda) \equiv \text{number of edges in } E \text{ not in } \widehat{E}^{\lambda}.$$

The oracle regularization level λ^* is then

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \{\mathrm{FP}(\lambda) + \mathrm{FN}(\lambda)\}.$$

The oracle score is $\mathrm{FP}(\lambda^*) + \mathrm{FN}(\lambda^*)$. Figure 6 shows boxplots of the oracle scores for the two methods, calculated using 100 simulations.

To illustrate the overall performance of these two methods over the full paths, ROC curves are shown in Figure 7, using

$$\left(1-\frac{\mathrm{FN}(\lambda)}{r}, 1-\frac{\mathrm{FP}(\lambda)}{\binom{p}{2}-r}\right).$$

The curves clearly show how the performance of both methods improves with sample size, and that the nonparanormal is superior to the Gaussian model in most cases.

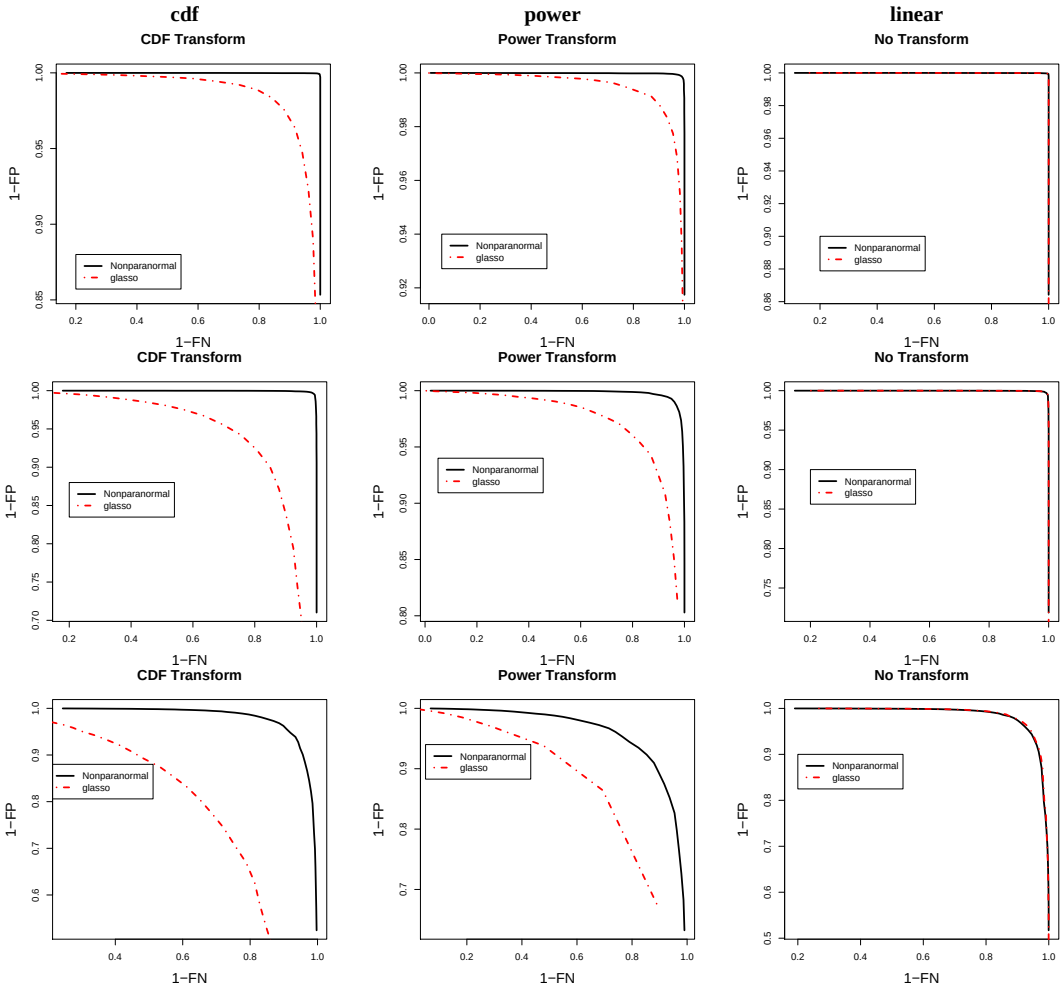


Figure 7: ROC curves for sample sizes $n = 1000, 500, 200$ (top, middle, bottom).

Let $\mathrm{FPE} \equiv \mathrm{FP}(\lambda^*)$ and $\mathrm{FNE} \equiv \mathrm{FN}(\lambda^*)$, Tables 1, 2, and 3 provide numerical comparisons of both methods on data sets with different transformations, where we repeat the experiments 100 times and report the average FPE and FNE values with the corresponding standard deviations. It's clear from the tables that the nonparanormal achieves significantly smaller errors than the glasso if the true distribution of the data is not multivariate Gaussian and achieves performance comparable to the glasso when the true distribution is exactly multivariate Gaussian.

Figure 8 shows typical runs for the cdf and power transformations. It's clear that when the glasso estimates the graph incorrectly, the mistakes include both false positives and negatives.

	Nonparanormal				glasso			
n	FPE	(sd(FPE))	FNE	(sd(FNE))	FPE	(sd(FPE))	FNE	(sd(FNE))
1000	0.10	(0.3333)	0.05	(0.2190)	3.73	(2.3904)	7.24	(3.2910)
900	0.18	(0.5389)	0.16	(0.4197)	3.31	(2.4358)	8.94	(3.2808)
800	0.16	(0.5069)	0.23	(0.5659)	3.80	(2.9439)	9.91	(3.4789)
700	0.26	(0.6295)	0.43	(0.7420)	3.45	(2.5519)	12.26	(3.5862)
600	0.33	(0.6039)	0.41	(0.6371)	3.31	(2.8804)	14.25	(4.0735)
500	0.58	(0.9658)	1.10	(1.0396)	3.18	(2.9211)	17.54	(4.4368)
400	0.71	(1.0569)	1.52	(1.2016)	1.58	(2.3535)	21.18	(4.9855)
300	1.37	(1.4470)	2.97	(2.0123)	0.67	(1.6940)	23.14	(5.0232)
200	2.03	(1.9356)	7.13	(3.4514)	0.01	(0.1000)	24.03	(4.9816)

Table 1: Quantitative comparison on the data set using the cdf transformation. For both FPE and FNE, the nonparanormal performs much better in general.

	Nonparanormal				glasso			
n	FPE	(sd(FPE))	FNE	(sd(FNE))	FPE	(sd(FPE))	FNE	(sd(FNE))
1000	0.27	(0.7086)	0.35	(0.6571)	2.89	(1.9482)	4.97	(2.9213)
900	0.38	(0.6783)	0.41	(0.6210)	2.98	(2.3697)	5.99	(3.0467)
800	0.25	(0.5751)	0.73	(0.8270)	4.10	(2.7834)	6.39	(3.3571)
700	0.69	(0.9067)	0.90	(1.0200)	4.42	(2.8891)	8.80	(3.9848)
600	0.92	(1.2282)	1.59	(1.5314)	4.64	(3.3830)	10.58	(4.2168)
500	1.17	(1.3413)	2.56	(2.3325)	4.00	(2.9644)	13.09	(4.4903)
400	1.88	(1.6470)	4.97	(2.7687)	3.14	(3.4699)	17.87	(4.7750)
300	2.97	(2.4181)	7.85	(3.5572)	1.36	(2.3805)	21.24	(4.7505)
200	2.82	(2.6184)	14.53	(4.3378)	0.37	(0.9914)	24.01	(5.0940)

Table 2: Quantitative comparison on the data set using the power transformation. For both FPE and FNE, the nonparanormal performs much better in general.

6.1.4 COMPARISON IN THE GAUSSIAN CASE

The previous experiments indicate that the nonparanormal works almost as well as the glasso in the Gaussian case. This initially appears surprising, since a parametric method is expected to be more efficient than a nonparametric method if the parametric assumption is correct. To manifest this efficiency loss, we conducted some experiments with very small n and relatively large p. For multivariate Gaussian models, Figure 9 shows results with $(n, p, s) = (50, 40, 1/8), (50, 100, 1/15)$

	Nonparanormal				glasso			
n	FPE	(sd(FPE))	FNE	(sd(FNE))	FPE	(sd(FPE))	FNE	(sd(FNE))
1000	0.10	(0.3333)	0.05	(0.2190)	0.09	(0.3208)	0.06	(0.2386)
900	0.24	(0.7537)	0.14	(0.4025)	0.22	(0.6447)	0.15	(0.4113)
800	0.17	(0.4277)	0.16	(0.3949)	0.16	(0.4431)	0.19	(0.4191)
700	0.25	(0.6871)	0.33	(0.8534)	0.29	(0.8201)	0.27	(0.7501)
600	0.37	(0.7740)	0.36	(0.7456)	0.36	(0.7722)	0.37	(0.6459)
500	0.28	(0.5874)	0.46	(0.7442)	0.25	(0.5573)	0.45	(0.6571)
400	0.55	(0.8453)	1.37	(1.2605)	0.47	(0.7713)	1.35	(1.2502)
300	1.24	(1.3715)	3.07	(1.7306)	0.98	(1.2058)	3.04	(1.8905)
200	1.62	(1.7219)	5.89	(2.7373)	1.55	(1.6779)	5.62	(2.6620)

Table 3: Quantitative comparison on the data set without any transformation. The two methods behave similarly, the glasso is slightly better.

and $(30, 100, 1/15)$. From the mean ROC curves, we see that nonparanormal does indeed behave worse than the glasso, suggesting some efficiency loss. However, from the corresponding boxplots, the efficiency reduction is relatively insignificant.

6.1.5 THE CASE WHEN $p \gg n$

Figure 10 shows results from a simulation of the nonparanormal using cdf transformations with $n = 200, p = 500$ and sparsity level $s = 1/40$. The boxplot shows that the nonparanormal outperforms the glasso. A typical run of the regularization paths confirms this conclusion, showing that the nonparanormal path separates the relevant and irrelevant dimensions very well. In contrast, with the glasso the relevant variables are "buried" among the irrelevant variables.

6.2 Gene Microarray Data

In this study, we consider a data set based on Affymetrix GeneChip microarrays for the plant *Arabidopsis thaliana*, (Wille et al., 2004). The sample size is $n = 118$. The expression levels for each chip are pre-processed by log-transformation and standardization. A subset of 40 genes from the isoprenoid pathway are chosen, and we study the associations among them using both the paranormal and nonparanormal models. Even though these data are generally treated as multivariate Gaussian in the previous analysis (Wille et al., 2004), our study shows that the results of the nonparanormal and the glasso are very different over a wide range of regularization parameters. This suggests the nonparanormal could support different scientific conclusions.

6.2.1 COMPARISON OF THE REGULARIZATION PATHS

We first compare the regularization paths of the two methods, in Figure 11. To generate the paths, we select 50 regularization parameters on an evenly spaced grid in the interval $[0.16, 1.2]$. Although

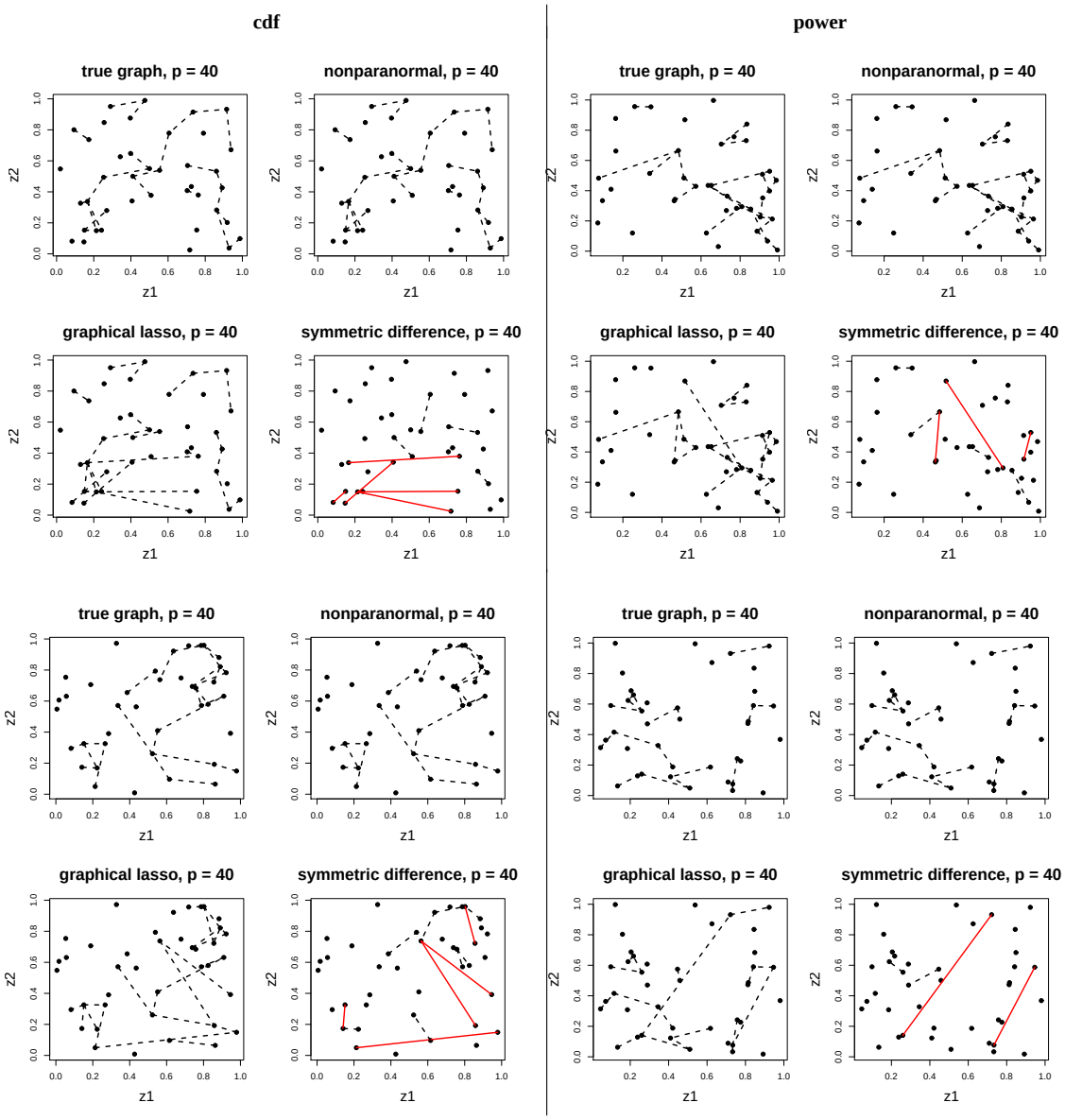


Figure 8: Typical runs for the two methods for $n = 1000$ using the cdf and power transformations. The dashed (black) lines in the symmetric difference plots indicate edges found by the glasso but not the nonparanormal, and vice-versa for the solid (red) lines.

the paths for the two methods look similar, there are some subtle differences. In particular, variables become nonzero in a different order, especially when the regularization parameter is in the range $\lambda \in [0.2, 0.3]$. As shown below, these subtle differences in the paths lead to different model selection behaviors.

6.2.2 COMPARISON OF THE ESTIMATED GRAPHS

Figure 12 compares the estimated graphs for the two methods at several values of the regularization parameter λ in the range $[0.16, 0.37]$. For each λ, we show the estimated graph from the nonparanormal in the first column. In the second column we show the graph obtained by scanning the full

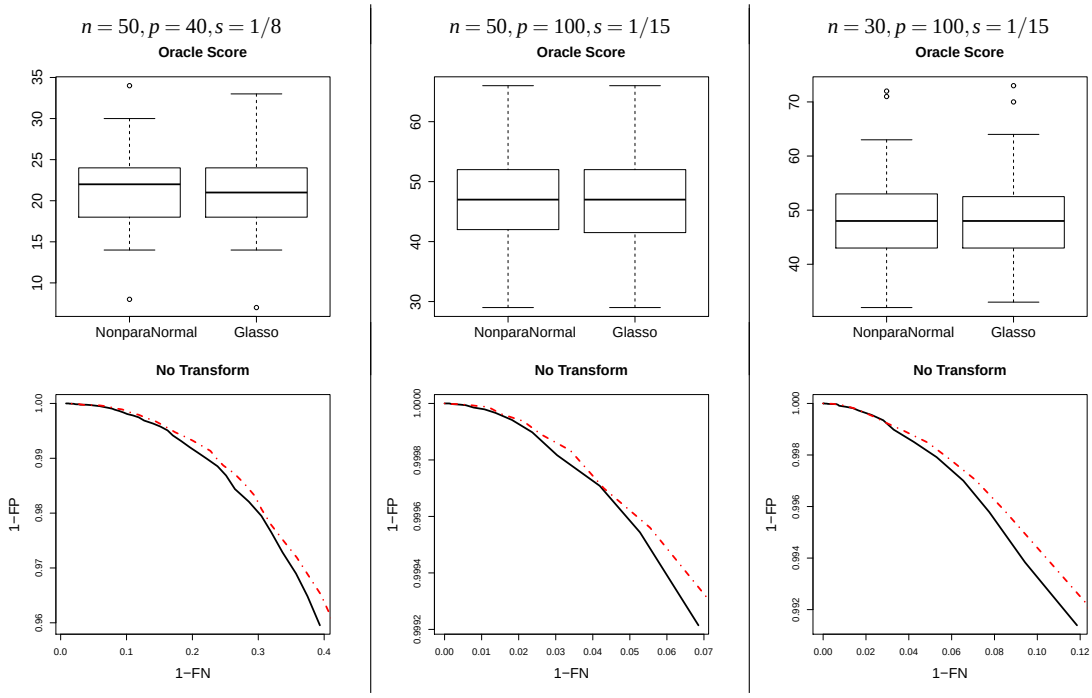


Figure 9: For Gaussian models, comparison of boxplots of the oracle scores and ROC curves for small n and relatively large p. The ROC curves suggest some efficiency loss of the nonparanormal; however, the corresponding boxplots indicate this loss is insignificant.

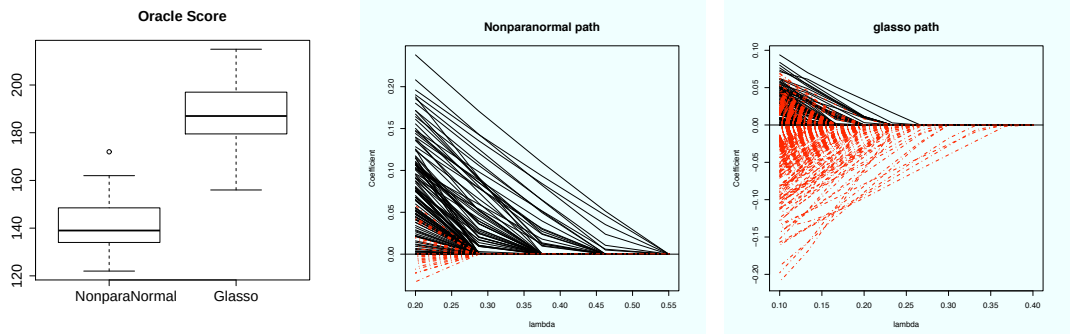


Figure 10: For the cdf transformation with $n = 200, p = 500, s = 1/40$, comparison of the boxplots and a typical run of the regularization paths. The nonparanormal paths separate the relevant from the irrelevant dimensions well. For the glasso, the relevant variables are "buried" in irrelevant variables.

regularization path of the glasso fit and finding the graph having the smallest symmetric difference with the nonparanormal graph. The symmetric difference graph is shown in in the third column. The closest glasso fit is different, with edges selected by the glasso not selected by the nonparanormal, and vice-versa. Several estimated transformations are plotted in Figure 13, which are are nonlinear. Interestingly, several of the differences between the fitted graphs are related to these variables.

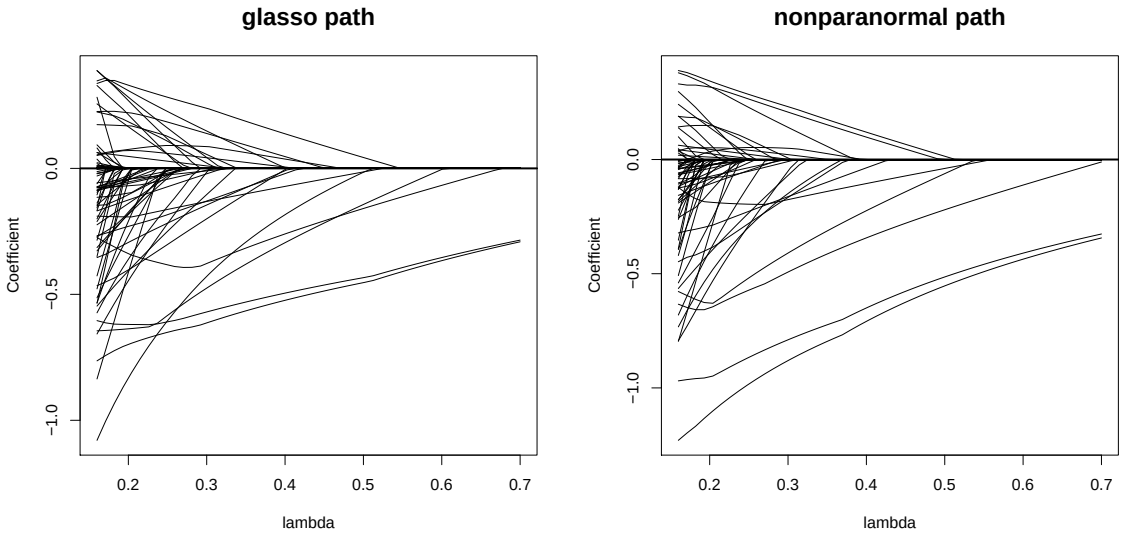


Figure 11: The regularization paths of both methods on the microarray data set. Although the paths for the two methods look similar, there are some subtle differences.

7. Proofs

We assume, without loss of generality from Lemma 3, that $\mu_j = 0$ and $\sigma_j = 1$ for all $j = 1, \ldots, p$. Thus, define $\widetilde{f}_j(x) \equiv \Phi^{-1}(\widetilde{F}_j(x))$ and $f_j(x) \equiv \Phi^{-1}(F_j(x))$, and let $g_j \equiv f_j^{-1}$.

7.1 Proof of Theorem 4

We start with some useful lemmas; the first is from Abramovich et al. (2006).

Lemma 11 *(Gaussian Distribution function vs. Quantile function) Let Φ and ϕ denote the distribution and density functions of a standard Gaussian random variable. Then*

$$\frac{\phi(t)}{2t} \leq 1 - \Phi(t) \leq \frac{\phi(t)}{t} \quad \text{if } t \geq 1$$

and

$$(\Phi^{-1})'(\eta) = \frac{1}{\phi(\Phi^{-1}(\eta))}.$$

Also, for $\eta \geq 0.99$, we have

$$\Phi^{-1}(\eta) = \sqrt{2\log\left(\frac{1}{1-\eta}\right)} - r(\eta) \tag{11}$$

where $r(\eta) \in [0, 1.5]$.

Lemma 12 *(Distribution function of the transformed random variable) For any $\alpha \in (-\infty, \infty)$*

$$\Phi^{-1}\left(F_j\left(g_j(\alpha\sqrt{\log n})\right)\right) = \alpha\sqrt{\log n}.$$

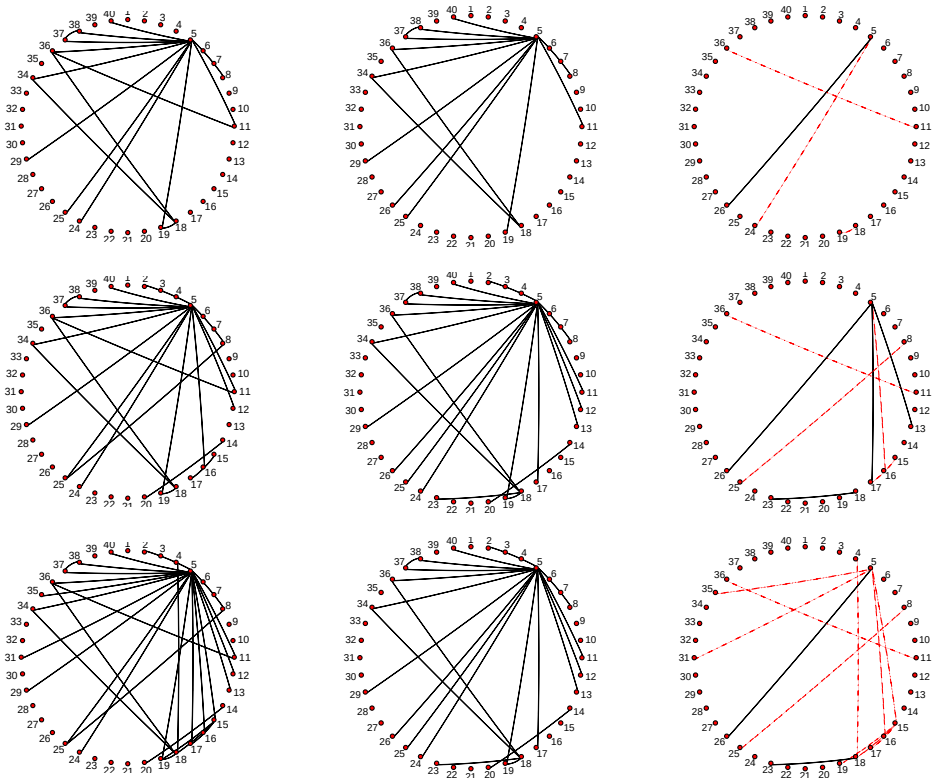


Figure 12: The nonparanormal estimated graph for three values of $\lambda = 0.2448, 0.2661, 0.30857$ (left column), the closest glasso estimated graph from the full path (middle) and the symmetric difference graph (right).

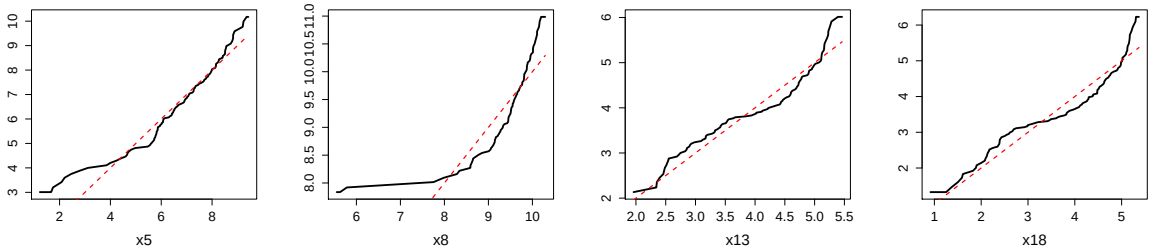


Figure 13: Estimated transformations for the microarray data set, indicating non-Gaussian marginals. The corresponding genes are among the nodes appearing in the symmetric difference graphs above.

Proof The statement follows from

$$F_j(t) = \mathbb{P}(X_j \leq t) = \mathbb{P}(g_j(Z_j) \leq t) = \mathbb{P}(Z_j \leq g_j^{-1}(t)) = \Phi\left(g_j^{-1}(t)\right). \tag{12}$$

which holds for any t. ■

Lemma 13 *(Gaussian maximal inequality) Let $W_1, \ldots, W_n$ be identically distributed standard Gaussian random variables (do not have to be independent). Then for any $\alpha > 0$*

$$\mathbb{P}\left(\max_{1\le i\le n} W_i > \sqrt{\alpha \log n}\right) \le \frac{1}{n^{\alpha/2-1}\sqrt{2\pi\alpha\log n}}.$$

Proof Using Mill's inequality, we have

$$\mathbb{P}\left(\max_{1\le i\le n} W_i > \sqrt{\alpha\log n}\right) \le \sum_{i=1}^{n} \mathbb{P}\left(W_i > \sqrt{\alpha\log n}\right) \le n\frac{\phi(\sqrt{\alpha\log n})}{\sqrt{\alpha\log n}} = \frac{1}{n^{\alpha/2-1}\sqrt{2\pi\alpha\log n}},$$

from which the result follows. ■

Lemma 14 *For any $\alpha > 0$ that satisfies $1-\delta_n - \Phi\left(\sqrt{\alpha\log n}\right) > 0$ for all n, we have*

$$\mathbb{P}\left[\widehat{F}_j\left(g_j\left(\sqrt{\alpha\log n}\right)\right) > 1-\delta_n\right] \le \exp\left\{-2n\left(1-\delta_n-\Phi\left(\sqrt{\alpha\log n}\right)\right)^2\right\}. \tag{13}$$

and

$$\mathbb{P}\left[\widehat{F}_j\left(g_j\left(-\sqrt{\alpha\log n}\right)\right) < \delta_n\right] \le \exp\left\{-2n\left(1-\delta_n-\Phi\left(\sqrt{\alpha\log n}\right)\right)^2\right\}. \tag{14}$$

Proof Using Hoeffding's inequality,

$$\begin{aligned}
&\mathbb{P}\left[\widehat{F}_j\left(g_j\left(\sqrt{\alpha\log n}\right)\right) > 1-\delta_n\right] \\
&= \mathbb{P}\left[\widehat{F}_j\left(g_j\left(\sqrt{\alpha\log n}\right)\right) - F_j\left(g_j\left(\sqrt{\alpha\log n}\right)\right) > 1-\delta_n - F_j\left(g_j\left(\sqrt{\alpha\log n}\right)\right)\right] \\
&\le \exp\left\{-2n\left(1-\delta_n-F_j\left(g_j\left(\sqrt{\alpha\log n}\right)\right)\right)^2\right\}.
\end{aligned}$$

Equation (13) then follows from equation (12). The proof of equation (14) uses the same argument. ■

Now let $M > 2$ and set $\beta = \frac{1}{2}$. We split the interval

$$\left[g_j(-\sqrt{M\log n}), g_j(\sqrt{M\log n})\right]$$

into two parts, the middle

$$\mathcal{M}_n \equiv \left(g_j\left(-\sqrt{\beta\log n}\right), g_j\left(\sqrt{\beta\log n}\right)\right)$$

and ends

$$\mathcal{E}_n \equiv \left[g_j\left(-\sqrt{M\log n}\right), g_j\left(-\sqrt{\beta\log n}\right)\right] \cup \left[g_j\left(\sqrt{\beta\log n}\right), g_j\left(\sqrt{M\log n}\right)\right].$$

The behaviors of the function estimates in these two regions are different, so we first establish bounds on the probability that a sample can fall in the end region $\mathcal{E}_n$.

Lemma 15 *Let* $A \equiv \sqrt{\frac{2}{\pi}}(\sqrt{M} - \sqrt{\beta})$. *Then*

$$\mathbb{P}(X_{1j} \in \mathcal{E}_n) \leq A\sqrt{\frac{\log n}{n^\beta}}, \ \ \forall j \in \{1,\ldots,p\}.$$

Proof Using Equation (12) and the mean value theorem, we have

$$\begin{aligned}
&\mathbb{P}(X_{1j} \in \mathcal{E}_n) \\
&= \mathbb{P}\left(X_{1j} \in \left[g_j(\sqrt{\beta\log n}), g_j(\sqrt{M\log n})\right]\right) + \mathbb{P}\left(X_{1j} \in \left[g_j(-\sqrt{M\log n}), g_j(-\sqrt{\beta\log n})\right]\right) \\
&= F_j\left(g_j(\sqrt{M\log n})\right) - F_j\left(g_j(\sqrt{\beta\log n})\right) + F_j\left(g_j(-\sqrt{\beta\log n})\right) - F_j\left(g_j(-\sqrt{M\log n})\right) \\
&= 2\left(\Phi(\sqrt{M\log n}) - \Phi(\sqrt{\beta\log n})\right) \\
&\leq 2\phi\left(\sqrt{\beta\log n}\right)\left(\sqrt{M\log n} - \sqrt{\beta\log n}\right).
\end{aligned}$$

The result of the lemma follows directly. ■

We next bound the error of the Winsorized estimate of a component function over the end region.

Lemma 16 *For all n, we have*

$$\sup_{t\in\mathcal{E}_n}\left|\Phi^{-1}(\widetilde{F}_j(t)) - \Phi^{-1}(F_j(t))\right| < \sqrt{2(M+2)\log n}, \ \ \forall j \in \{1,\ldots,p\}.$$

Proof From Lemma 12 and the definition of $\mathcal{E}_n$, we have

$$\sup_{t\in\mathcal{E}_n}\left|\Phi^{-1}(F_j(t))\right| \in \left[0, \sqrt{M\log n}\right].$$

Given the fact that $\delta_n = \dfrac{1}{4n^{1/4}\sqrt{\pi\log n}}$, we have $\widetilde{F}_j(t) \in \left(\dfrac{1}{n}, 1-\dfrac{1}{n}\right)$. Therefore, from Equation (11),

$$\sup_{t\in\mathcal{E}_n}\left|\Phi^{-1}\left(\widetilde{F}_j(t)\right)\right| \in \left[0, \sqrt{2\log n}\right).$$

The result follows from the triangle inequality and $\sqrt{M} + \sqrt{2} \leq \sqrt{2(M+2)}$. ■

Now for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{j,k}\left|S_n(\widetilde{f})_{jk} - S_n(f)_{jk}\right| > 2\varepsilon\right) \\
&= \mathbb{P}\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n\left\{\widetilde{f}_j(X_{ij})\widetilde{f}_k(X_{ik}) - f_j(X_{ij})f_k(X_{ik}) - \mu_n(\widetilde{f}_j)\mu_n(\widetilde{f}_k) + \mu_n(f_j)\mu_n(f_k)\right\}\right| > 2\varepsilon\right) \\
&\leq \mathbb{P}\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n\left(\widetilde{f}_j(X_{ij})\widetilde{f}_k(X_{ik}) - f_j(X_{ij})f_k(X_{ik})\right)\right| > \varepsilon\right) \\
&\quad + \mathbb{P}\left(\max_{j,k}\left|\mu_n(\widetilde{f}_j)\mu_n(\widetilde{f}_k) - \mu_n(f_j)\mu_n(f_k)\right| > \varepsilon\right).
\end{aligned}$$

We only need to analyze the rate for the first term above, since the second one is of higher order (Cai et al., 2008). Let

$$\Delta_i(j,k) \equiv \widetilde{f}_j(X_{ij})\,\widetilde{f}_k(X_{ik}) - f_j(X_{ij})\,f_k(X_{ik})$$

and

$$\Theta_{t,s}(j,k) \equiv \widetilde{f}_j(t)\,\widetilde{f}_k(s) - f_j(t)\,f_k(s).$$

We define the event $\mathcal{A}_n$ as

$$\mathcal{A}_n \equiv \left\{ g_j\left(-\sqrt{M\log n}\right) \le X_{1j},\ldots,X_{nj} \le g_j\left(\sqrt{M\log n}\right), j=1,\ldots,p \right\}.$$

Then, by Lemma 13, when $M \ge 2(\xi+1)$, we have

$$\mathbb{P}(\mathcal{A}_n^c) \le \mathbb{P}\left(\max_{i,j\in\{1,\ldots,n\}\times\{1,\ldots,p\}} |f_j(X_{ij})| > \sqrt{2\log(np)} \right) \le \frac{1}{2\sqrt{\pi\log(np)}}.$$

Therefore

$$\begin{aligned}
\mathbb{P}\left(\max_{j,k} \left| \frac{1}{n}\sum_{i=1}^n \Delta_i(j,k) \right| > \varepsilon \right) &\le \mathbb{P}\left(\max_{j,k} \left| \frac{1}{n}\sum_{i=1}^n \Delta_i(j,k) \right| > \varepsilon, \mathcal{A}_n \right) \quad + \quad \mathbb{P}(\mathcal{A}_n^c) \\
&\le \mathbb{P}\left(\max_{j,k} \left| \frac{1}{n}\sum_{i=1}^n \Delta_i(j,k) \right| > \varepsilon, \mathcal{A}_n \right) \quad + \quad \frac{1}{2\sqrt{\pi\log(np)}}.
\end{aligned}$$

Thus, we only need to carry out our analysis on the event $\mathcal{A}_n$. On this event, we have the following decomposition:

$$\begin{aligned}
&\mathbb{P}\left(\max_{j,k} \left| \frac{1}{n}\sum_{i=1}^n \Delta_i(j,k) \right| > \varepsilon,\ \mathcal{A}_n \right) \\
&\le \quad \mathbb{P}\left(\max_{j,k} \frac{1}{n} \sum_{X_{ij}\in\mathcal{M}_n, X_{ik}\in\mathcal{M}_n} |\Delta_i(j,k)| > \frac{\varepsilon}{4} \right) + \mathbb{P}\left(\max_{j,k} \frac{1}{n} \sum_{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n} |\Delta_i(j,k)| > \frac{\varepsilon}{4} \right) \\
&\quad + 2\mathbb{P}\left(\max_{j,k} \frac{1}{n} \sum_{X_{ij}\in\mathcal{M}_n, X_{ik}\in\mathcal{E}_n} |\Delta_i(j,k)| > \frac{\varepsilon}{4} \right).
\end{aligned}$$

We now analyze each of these terms separately.

Lemma 17 *On the event $\mathcal{A}_n$, let $\beta = 1/2$ and $\varepsilon \ge C_M \sqrt{\dfrac{\log p \log^2 n}{n^{1/2}}}$, then*

$$\mathbb{P}\left(\max_{j,k} \frac{1}{n} \sum_{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n} |\Delta_i(j,k)| > \frac{\varepsilon}{4} \right) = o(1).$$

Proof We define

$$\theta_1 \equiv \frac{n^{\beta/2}\varepsilon}{8A\sqrt{\log n}}$$

with the same parameter A as in Lemma 15. Such a θ_1 guarantees that

$$\frac{n\varepsilon}{4\theta_1} - nA\sqrt{\frac{\log n}{n^\beta}} = nA\sqrt{\frac{\log n}{n^\beta}} > 0.$$

By Lemma 15, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n\}} > \frac{\varepsilon}{4\theta_1}\right) &\le \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}_{\{X_{ij}\in\mathcal{E}_n\}} > \frac{n\varepsilon}{4\theta_1}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n \left(\mathbf{1}_{\{X_{ij}\in\mathcal{E}_n\}} - \mathbb{P}(X_{1j}\in\mathcal{E}_n)\right) > \frac{n\varepsilon}{4\theta_1} - n\mathbb{P}(X_{1j}\in\mathcal{E}_n)\right) \\
&\le \mathbb{P}\left(\sum_{i=1}^n \left(\mathbf{1}_{\{X_{ij}\in\mathcal{E}_n\}} - \mathbb{P}(X_{1j}\in\mathcal{E}_n)\right) > \frac{n\varepsilon}{4\theta_1} - nA\sqrt{\frac{\log n}{n^\beta}}\right).
\end{aligned}$$

Using the Bernstein's inequality, for $\beta = \frac{1}{2}$,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n\}} > \frac{\varepsilon}{4\theta_1}\right) &\le \mathbb{P}\left(\sum_{i=1}^n \left(\mathbf{1}_{\{X_{ij}\in\mathcal{E}_n\}} - \mathbb{P}(X_{1j}\in\mathcal{E}_n)\right) > nA\sqrt{\frac{\log n}{n^\beta}}\right) \\
&\le \exp\left(-\frac{c_1 n^{2-\beta}\log n}{c_2 n^{1-\beta/2}\sqrt{\log n} + c_3 n^{1-\beta/2}\sqrt{\log n}}\right) = o(1),
\end{aligned}$$

where $c_1, c_2, c_3 > 0$ are generic constants.

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n}|\Delta_i(j,k)| > \frac{\varepsilon}{4}\right) \\
&= \mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n}|\Delta_i(j,k)| > \frac{\varepsilon}{4}, \max_{j,k}\sup_{t\in\mathcal{E}_n, s\in\mathcal{E}_n}|\Theta_{t,s}(j,k)| > \theta_1\right) \\
&\quad + \mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n}|\Delta_i(j,k)| > \frac{\varepsilon}{4}, \max_{j,k}\sup_{t\in\mathcal{E}_n, s\in\mathcal{E}_n}|\Theta_{t,s}(j,k)| \le \theta_1\right) \\
&\le \mathbb{P}\left(\max_{j,k}\sup_{t\in\mathcal{E}_n, s\in\mathcal{E}_n}|\Theta_{t,s}(j,k)| > \theta_1\right) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_{ij}\in\mathcal{E}_n, X_{ik}\in\mathcal{E}_n\}} > \frac{\varepsilon}{4\theta_1}\right) \\
&= \mathbb{P}\left(\max_{j,k}\sup_{t\in\mathcal{E}_n, s\in\mathcal{E}_n}|\Theta_{t,s}(j,k)| > \theta_1\right) + o(1).
\end{aligned}$$

Now, we analyze the first term

$$\begin{aligned}\mathbb{P}\left(\max_{j,k}\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|\Theta_{t,s}(j,k)|>\theta_1\right) &\leq p^2\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|\Theta_{t,s}(j,k)|>\theta_1\right)\\ &= p^2\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|\widetilde{f}_j(t)\widetilde{f}_k(s)-f_j(t)f_k(s)|>\theta_1\right).\end{aligned}$$

By adding and subtracting terms $f_j(t)$ and $f_s(t)$, we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|\widetilde{f}_j(t)\widetilde{f}_k(s)-f_j(t)f_k(s)|>\theta_1\right)\\ &\qquad\leq \mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|(\widetilde{f}_j(t)-f_j(t))(\widetilde{f}_k(s)-f_k(s))|>\frac{\theta_1}{3}\right)\\ &\qquad\quad+\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|(\widetilde{f}_j(t)-f_j(t))|\cdot|f_k(s)|>\frac{\theta_1}{3}\right)\\ &\qquad\quad+\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|(\widetilde{f}_k(s)-f_k(s))|\cdot|f_j(t)|>\frac{\theta_1}{3}\right).\end{aligned}$$

The first term can further be decomposed to be

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|(\widetilde{f}_j(t)-f_j(t))(\widetilde{f}_k(s)-f_k(s))|>\frac{\theta_1}{3}\right)\\ &\quad\leq \mathbb{P}\left(\sup_{t\in\mathcal{E}_n}|(\widetilde{f}_j(t)-f_j(t))|>\sqrt{\frac{\theta_1}{3}}\right)+\mathbb{P}\left(\sup_{s\in\mathcal{E}_n}|(\widetilde{f}_k(s)-f_k(s))|>\sqrt{\frac{\theta_1}{3}}\right).\end{aligned}$$

Also, from the definition of $\mathcal{E}_n$, we have

$$\sup_{t\in\mathcal{E}_n}|f_j(t)|=\sup_{t\in\mathcal{E}_n}\left|g_j^{-1}(t)\right|\leq\sqrt{M\log n}.$$

Since $\varepsilon\geq C_M\sqrt{\frac{\log p\log^2 n}{n^{1/2}}}$, we have

$$\frac{\theta_1}{3}=\frac{n^{\beta/2}\varepsilon}{24A\sqrt{\log n}}\geq\frac{C_M\sqrt{\log p\log^2 n}}{24A\sqrt{\log n}}=2(M+2)\log n.$$

This implies that

$$\sqrt{\frac{\theta_1}{3}}\geq\sqrt{2(M+2)\log n}\ \text{ and }\ \frac{\theta_1}{3\sqrt{M\log n}}\geq\sqrt{2(M+2)\log n}.$$

Then, from Lemma 16, we get

$$\mathbb{P}\left(\sup_{t\in\mathcal{E}_n}|(\widetilde{f}_j(t)-f_j(t))|>\sqrt{\frac{\theta_1}{3}}\right)=0$$

and

$$\mathbb{P}\left(\sup_{t\in\mathcal{E}_n,s\in\mathcal{E}_n}|(\widetilde{f}_j(t)-f_j(t))|\cdot|f_k(s)|>\frac{\theta_1}{3}\right)=0.$$

The claim of the lemma then follows directly. ■

Remark 18 *From the above analysis, we see that the data in the tails doesn't affect the rate. Using exactly the same argument, we can also show that*

$$\mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{M}_n,X_{ik}\in\mathcal{E}_n}|\Delta_i(j,k)|>\frac{\varepsilon}{4}\right)=o(1).$$

Lemma 19 *On the event $\mathcal{A}_n$, let $\beta=1/2$ and $\varepsilon\geq C_M\sqrt{\frac{\log p\log^2 n}{n^{1/2}}}$. We have*

$$\mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{M}_n,X_{ik}\in\mathcal{M}_n}|\Delta_i(j,k)|>\frac{\varepsilon}{4}\right)\leq 2\exp\left(2\log p-\frac{n^{1/2}\varepsilon^2}{1232\pi^2\log^2 n}\right)+2\exp\left(2\log p-\frac{n^{1/2}}{8\pi\log n}\right).$$

Proof We have

$$\begin{aligned}\mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{M}_n,X_{ik}\in\mathcal{M}_n}|\Delta_i(j,k)|>\frac{\varepsilon}{4}\right)&\leq p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n,s\in\mathcal{M}_n}|\widetilde{f}_j(t)\widetilde{f}_k(s)-f_j(t)f_k(s)|>\frac{\varepsilon}{4}\right)\\&\leq p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n,s\in\mathcal{M}_n}|(\widetilde{f}_j(t)-f_j(t))(\widetilde{f}_k(s)-f_k(s))|>\frac{\varepsilon}{12}\right)\\&\quad+2p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n,s\in\mathcal{M}_n}|(\widetilde{f}_j(t)-f_j(t))|\cdot|f_k(s)|>\frac{\varepsilon}{12}\right).\end{aligned}$$

Further, since

$$\sup_{t\in\mathcal{M}_n}|f_j(t)|=\sup_{t\in\mathcal{M}_n}\left|g_j^{-1}(t)\right|=\sqrt{\beta\log n}$$

and $\sup_{t\in\mathcal{M}_n,s\in\mathcal{M}_n}|(\widetilde{f}_j(t)-f_j(t))(\widetilde{f}_k(s)-f_k(s))|$ is of higher order than $\sup_{t\in\mathcal{M}_n,s\in\mathcal{M}_n}|(\widetilde{f}_j(t)-f_j(t))|\cdot|f_k(s)|$, we only need to analyze the term $\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}|(\widetilde{f}_j(t)-f_j(t))|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)$.

Since $\delta_n=\frac{1}{4n^{\beta/2}\sqrt{2\pi\beta\log n}}$, using Mill's inequality we have

$$2\delta_n=\frac{\phi(\sqrt{\beta\log n})}{2\sqrt{\beta\log n}}\leq 1-\Phi(\sqrt{\beta\log n}).$$

This implies that

$$1-\delta_n-\Phi(\sqrt{\beta\log n})\geq\delta_n>0.$$

Using Lemma 14, we have

$$p^2\mathbb{P}\left(\widehat{F}_j\left(g_j\left(\sqrt{\beta\log n}\right)\right)>1-\delta_n\right)\leq p^2\exp\left(-2n\delta_n^2\right)=\exp\left(2\log p-\frac{n^{1-\beta}}{(16\pi\beta\log n)}\right) \tag{15}$$

and

$$p^2\mathbb{P}\left(\widehat{F}_j\left(g_j\left(-\sqrt{\beta\log n}\right)\right)<\delta_n\right)\leq\exp\left(2\log p-\frac{n^{1-\beta}}{(16\pi\beta\log n)}\right). \tag{16}$$

Define an event $\mathcal{B}_n$ as

$$\mathcal{B}_n\equiv\left\{\delta_n\leq\widehat{F}_j\left(g_j\left(\sqrt{\beta\log n}\right)\right)\leq 1-\delta_n, j=1,\ldots,p\right\}.$$

From (15) and (16), it is easy to see that

$$\mathbb{P}(\mathcal{B}_n^c)\leq 2\exp\left(2\log p-\frac{n^{1/2}}{8\pi\log n}\right).$$

From the definition of $\widetilde{F}_j$, we have

$$\begin{aligned}
&p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}|\widetilde{f}_j(t)-f_j(t)|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)\\
&\leq\quad p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}\left|\Phi^{-1}\left(\widetilde{F}_j(t)\right)-\Phi^{-1}\left(F_j(t)\right)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}},\mathcal{B}_n\right)+\mathbb{P}(\mathcal{B}_n^c).\\
&\leq\quad p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}\left|\Phi^{-1}\left(\widehat{F}_j(t)\right)-\Phi^{-1}\left(F_j(t)\right)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)+2\exp\left(2\log p-\frac{n^{1/2}}{8\pi\log n}\right).
\end{aligned}$$

Define

$$T_{1n}\equiv\max\left\{F_j\left(g_j\left(\sqrt{\beta\log n}\right)\right),1-\delta_n\right\}\ \text{ and }\ T_{2n}\equiv 1-\min\left\{F_j\left(g_j\left(-\sqrt{\beta\log n}\right)\right),\delta_n\right\}.$$

From Equation (12) and the fact that $1-\delta_n\geq\Phi\left(\sqrt{\beta\log n}\right)$, we have that

$$T_{1n}=T_{2n}=1-\delta_n.$$

Thus, by the mean value theorem,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}\left|\Phi^{-1}\left(\widehat{F}_j(t)\right)-\Phi^{-1}\left(F_j(t)\right)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)\\
&\leq\quad\mathbb{P}\left((\Phi^{-1})'\left(\max\{T_{1n},T_{2n}\}\right)\sup_{t\in\mathcal{M}_n}\left|\widehat{F}_j(t)-F_j(t)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)\\
&=\quad\mathbb{P}\left((\Phi^{-1})'(1-\delta_n)\sup_{t\in\mathcal{M}_n}\left|\widehat{F}_j(t)-F_j(t)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right).
\end{aligned}$$

Finally, using the Dvoretzky-Kiefer-Wolfowitz inequality,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}\left|\Phi^{-1}\left(\widehat{F}_j(t)\right)-\Phi^{-1}\left(F_j(t)\right)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)\\
&\qquad\leq\ \mathbb{P}\left(\sup_{t\in\mathcal{M}_n}\left|\widehat{F}_j(t)-F_j(t)\right|>\frac{\varepsilon}{(\Phi^{-1})'(1-\delta_n)\,12\sqrt{\beta\log n}}\right)\\
&\qquad\leq\ 2\exp\left(-2\frac{n\varepsilon^2}{144\beta\log n\left[(\Phi^{-1})'(1-\delta_n)\right]^2}\right).
\end{aligned}$$

Furthermore, by Lemma 11,

$$(\Phi^{-1})'(1-\delta_n)=\frac{1}{\phi(\Phi^{-1}(1-\delta_n))}\leq\frac{1}{\phi\left(\sqrt{2\log\frac{1}{\delta_n}}\right)}=\sqrt{2\pi}\left(\frac{1}{\delta_n}\right)=8\pi n^{\beta/2}\sqrt{\beta\log n}.$$

This implies that

$$p^2\mathbb{P}\left(\sup_{t\in\mathcal{M}_n}\left|\Phi^{-1}\left(\widehat{F}_j(t)\right)-\Phi^{-1}\left(F_j(t)\right)\right|>\frac{\varepsilon}{12\sqrt{\beta\log n}}\right)\leq 2\exp\left(2\log p-\frac{n^{1/2}\varepsilon^2}{1232\pi^2\log^2 n}\right).$$

In summary, we have

$$\mathbb{P}\left(\max_{j,k}\frac{1}{n}\sum_{X_{ij}\in\mathcal{M}_n,X_{ik}\in\mathcal{E}_n}|\Delta_i(j,k)|>\frac{\varepsilon}{4}\right)\leq 2\exp\left(2\log p-\frac{n^{1/2}\varepsilon^2}{1232\pi^2\log^2 n}\right)+2\exp\left(2\log p-\frac{n^{1/2}}{8\pi\log n}\right)$$

This finish the proof. ■

The conclusion of Theorem 4 follows from Lemma 17 and Lemma 19.

7.2 Proof of Theorem 8

Proof First note that the population and sample risks are

$$\begin{aligned}
R(f,\Omega) &= \frac{1}{2}\left\{\mathrm{tr}\left[\Omega\mathbb{E}(f(X)f(X)^T\right]-\log|\Omega|-p\log(2\pi)\right\}\\
\widehat{R}(f,\Omega) &= \frac{1}{2}\left\{\mathrm{tr}\left[\Omega S_n(f)\right]-\log|\Omega|-p\log(2\pi)\right\}.
\end{aligned}$$

Therefore, for all $(f,\Omega)\in\mathcal{M}_n^p\oplus\mathcal{C}_n$, we have

$$\begin{aligned}
|\widehat{R}(f,\Omega)-R(f,\Omega)| &= \frac{1}{2}\left|\mathrm{tr}\left[\Omega\left(\mathbb{E}[ff^T]-S_n(f)\right)\right]\right|\\
&\leq \frac{1}{2}\|\Omega\|_1\max_{jk}\sup_{f_j,f_k\in\mathcal{M}_n}|\mathbb{E}(f_j(X_j)f_k(X_k)-S_n(f)_{jk}|\\
&\leq \frac{L_n}{2}\max_{jk}\sup_{f_j,f_k\in\mathcal{M}_n}|\mathbb{E}(f_j(X_j)f_k(X_k)-S_n(f)_{jk}|.
\end{aligned}$$

Now, if $\mathcal{F}$ is a class of functions, we have

$$\mathbb{E}\left(\sup_{g\in\mathcal{F}}|\widehat{\mu}(g)-\mu(g)|\right)\leq\frac{CJ_{[]}(\|F\|_\infty,\mathcal{F})}{\sqrt{n}} \tag{17}$$

for some $C>0$, where $F(x)=\sup_{g\in cF}|g(x)|$, $\mu(g)=\mathbb{E}(g(X))$ and $\widehat{\mu}(g)=n^{-1}\sum_{i=1}^n g(X_i)$ (see Corollary 19.35 of van der Vaart 1998). Here the bracketing integral is defined to be

$$J_{[]}(\delta,\mathcal{F})=\int_0^\delta\sqrt{\log N_{[]}(u,\mathcal{F})}\,du$$

where $\log N_{[]}(\varepsilon,\mathcal{F})$ is the bracketing entropy. For the class of one dimensional, bounded and monotone functions, the bracketing entropy satisfies

$$\log N_{[]}(\varepsilon,\mathcal{M})\leq K\left(\frac{1}{\varepsilon}\right)$$

for some $K>0$ (van der Vaart and Wellner, 1996).

Now, let $\mathcal{P}_{n,p}$ be the class of all functions of the form $m(x)=f_j(x_j)f_k(x_k)$ for $j,k\in\{1,\ldots,p\}$, where $f_j\in\mathcal{M}_n$ for each j. Then the bracketing entropy satisfies

$$\log N_{[]}(C\sqrt{\log n},\mathcal{P}_{n,p})\leq 2\log p+K\left(\frac{1}{\varepsilon}\right)$$

and the bracketing integral satisfies $J_{[]}(C\sqrt{\log n},\mathcal{P}_{n,p})=O(\sqrt{\log n\log p})$. It follows from (17) and Markov's inequality that

$$\max_{jk}\sup_{f_j,f_k\in\mathcal{M}_n}|S_n(f)_{jk}-\mathbb{E}(f_j(X_j)f_k(X_k)|=O_P\left(\sqrt{\frac{\log n\log p}{n}}\right)=O_P\left(\sqrt{\frac{\log n}{n^{1-\xi}}}\right).$$

Therefore,

$$\sup_{(f,\Omega)\in\mathcal{M}_n^p\oplus\mathcal{C}_n}|\widehat{R}(f,\Omega)-R(f,\Omega)|=O_P\left(\frac{L_n\sqrt{\log n}}{n^{(1-\xi)/2}}\right).$$

As a consequence, we have

$$\begin{aligned}
R(f^*,\Omega^*) &\leq R(\widetilde{f}_n,\widehat{\Omega}_n)\\
&\leq \widehat{R}(\widetilde{f}_n,\widehat{\Omega}_n)+O_P\left(\frac{L_n\sqrt{\log n}}{n^{(1-\xi)/2}}\right)\\
&\leq \widehat{R}(f^*,\Omega^*)+O_P\left(\frac{L_n\sqrt{\log n}}{n^{(1-\xi)/2}}\right)\\
&\leq R(f^*,\Omega^*)+O_P\left(\frac{L_n\sqrt{\log n}}{n^{(1-\xi)/2}}\right)
\end{aligned}$$

and the conclusion follows. ■

8. Concluding Remarks

In this paper we have introduced the nonparanormal, a type of Gaussian copula with nonparametric marginals that is suitable for estimating high dimensional undirected graphs. The nonparanormal can be viewed as an extension of sparse additive models to the setting of graphical models. We proposed an estimator for the component functions that is based on thresholding the tails of the empirical distribution function at appropriate levels. A theoretical analysis was given to bound the difference between the sample covariance with respect to these estimated functions and the true sample covariance. This analysis was leveraged with the recent work of Ravikumar et al. (2009b) and Rothman et al. (2008) to obtain consistency results for the nonparanormal. Computationally, fitting a high dimensional nonparanormal is no more difficult than estimating a multivariate Gaussian, and indeed one can exploit existing software for the graphical lasso. Our experimental results indicate that the sparse nonparanormal can give very different results than a sparse Gaussian graphical model. This suggests that it may be a useful tool for relaxing the normality assumption, which is often made only for convenience.

Acknowledgments

We thank Zoubin Ghahramani, Michael Jordan, and the anonymous reviewers for helpful comments on this work. The research reported here was supported in part by NSF grant CCF-0625879 and a grant from Google.

References

Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516, March 2008.

Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. Technical report, Wharton School, Statistics Department, University of Pennsylvania, 2008.

Mathias Drton and Michael D. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22(3):430–449, 2007.

Mathias Drton and Michael D. Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

Jerome H. Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2007.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Chapman & Hall Ltd., 1999.

Chris A. J. Klaassen and Jon A. Wellner. Efficient estimation in the bivariate normal copula model: Normal margins are least-favorable. *Bernoulli*, 3(1):55–77, 1997.

Colin L. Mallows, editor. *The collected works of John W. Tukey. Volume VI: More mathematical, 1938–1984*. Wadsworth & Brooks/Cole, 1990.

Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

Pradeep Ravikumar, Han Liu, John Lafferty, and Larry Wasserman. SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems 20*, pages 1201–1208. MIT Press, Cambridge, MA, 2008.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B, Methodological*, 2009a. To appear.

Pradeep Ravikumar, Martin Wainwright, Garvesh Raskutti, and Bin Yu. Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems 22*, Cambridge, MA, 2009b. MIT Press.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Abe Sklar. Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris 8*, pages 229–231, 1959.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological*, 58:267–288, 1996.

Hideatsu Tsukahara. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33:357–375, 2005.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, 1996.

Anja Wille et al. Sparse Gaussian graphical modelling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, 5:R92, 2004.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.