PsyArXiv
April 7, 2018

Back to the Basics: Rethinking Partial Correlation Network Methodology

Donald R. Williams
University of California, Davis

Philippe Rast
University of California, Davis

Gaussian graphical models are an increasingly popular technique in psychology to characterize relationships among observed variables. These relationships are represented as covariances in the precision matrix. Standardizing this covariance matrix and reversing the sign yields corresponding partial correlations that imply pairwise dependencies in which the effects of all other variables have been controlled for. In order to estimate the precision matrix, the graphical lasso (glasso) has emerged as the default estimation method, which uses ℓ_1-based regularization. Glasso was developed and optimized for high dimensional settings where the number of variables (p) exceeds the number of observations (n) which are uncommon in psychological applications. Here we propose to go "back to the basics", wherein the precision matrix is first estimated with non-regularized maximum likelihood and then Fisher Z-transformed confidence intervals are used to determine non-zero relationships. We first show the exact correspondence between the confidence level and specificity, which is due to 1 - specificity denoting the false positive rate (i.e., α). With simulations in low-dimensional settings ($p \ll n$), we then demonstrate superior performance compared to glasso for determining conditional relationships, in addition to frequentist risk measured with various loss functions. Further, our results indicate that glasso is inconsistent for the purpose of model selection, whereas the proposed method converged on the *true* model with a probability that approached 100%. We end by discussing implications for estimating Gaussian graphical models in psychology.

Keywords: Gaussian graphical model, maximum likelihood, Fisher Z-transformation, partial correlation, confidence interval, ℓ_1 regularization

Introduction

An important goal for psychological science is developing methods to characterize relationships between variables. The customary approach uses structural equation models (SEM) to connect latent factors on a structural level to a number of observed measurements (Maccallum & Austin, 2000). More recently, Gaussian graphical models (GGMs) have been proposed as an alternative approach for describing the relation among variables, and they have become increasingly popular in psychology (Borsboom & Cramer, 2013; Epskamp & Fried, 2016; Van Borkulo et al., 2014). Rather than assessing a hypothesized model structure, as in a SEM, GGMs seek to capture conditional relationships (i.e., direct effects) between a set of observed variables. On the computational level, this is accomplished by identifying non-zero covariances in the off-diagonal elements of the inverse-covariance matrix (i.e., precision matrix) (Dempster, 1972). When these covariances are standardized and the sign reversed, they correspond to partial correlations that imply pairwise dependencies in which the linear effects of all other variables have been controlled for (Fan, Liao, & Liu, 2016; Whittaker, 1990). Since direct effects allow for rich inferences, this has resulted in a growing body of literature called "network modeling" in both methodological (Epskamp & Fried, 2016; Epskamp, Kruis, & Marsman, 2017) and applied contexts (McNally et al., 2015; Rhemtulla et al., 2016).

The default approach for estimating network models in psychology uses ℓ_1 regularization (e.g., a form of penalized maximum likelihood) (Epskamp & Fried, 2016), which can simultaneously improve predictive accuracy and perform variable selection by reducing some parameters to exactly zero (Dalalyan, Hebiri, & Lederer, 2017). In the context of regression, this is known as the lasso (least absolute shrinkage and selector operator) method (Dalalyan et al., 2017), whereas the extension to multivariate settings is called the graphical

Research reported in this publication was supported by three funding sources: (1) The National Academies of Sciences, Engineering, and Medicine FORD foundation pre-doctoral fellowship to DRW; (2) The National Science Foundation Graduate Research Fellowship to DRW; and (3) the National Institute On Aging of the National Institutes of Health under Award Number R01AG050720 to PR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Academies of Sciences, Engineering, and Medicine, the National Science Foundation, or the National Institutes of Health.

lasso (glasso) (Friedman, Hastie, & Tibshirani, 2008). Importantly, the glasso method was primarily developed to overcome challenges in high-dimensional settings, in which the number of variables (p) often exceeds the number of observations (n) (Fan et al., 2016). In these situations, the covariance matrix cannot be inverted due to singularity (Hartlap, Simon, & Schneider, 2007), which is overcome by the glasso method. Accordingly, most of the simulation work has focused on high-dimensional settings ($n < p$), where model selection consistency is not typically evaluated in more common asymptotic settings ($n \to \infty$) (Ha & Sun, 2014; Heinävaara, Leppä-aho, Corander, & Honkela, 2016; Peng, Wang, Zhou, & Zhu, 2009). Further, in behavioral science applications, the majority of network models are fit in low-dimensional settings ($p \ll n$) (Costantini et al., 2015; Rhemtulla et al., 2016). Unfortunately, model selection consistency has not been demonstrated with simulation studies representative of typical psychological applications. One aim of the current work is to fill this gap by investigating the properties of the most common glasso estimation techniques in situations where p is representative of the psychological literature and fixed, while n increases. This has a straightforward translation to applied settings: when a psychometric scale has been decided on (the number of variables p is fixed), an important goal is obtaining the smallest possible sample (n) to accurately estimate the network. A consistent method for model selection will ideally converge on the *true* model, with a probability approaching 100%, at some point as the sample size becomes larger (Casella, Girón, Martinez, & Moreno, 2009, $n \to \infty$).

There is some indication in the literature that the performance ℓ_1 regularization does not generalize to all settings. Especially in the context of graphical models. For example, Heinävaara et al. (2016) demonstrated that ℓ_1-based methods have sub-optimal performance with highly correlated variables, and that the assumptions for consistency are rarely met in their particular field of study (genomics). According to Heinävaara et al. (2016):

> Our results strongly suggest that users of the numerous ℓ_1-penalised and other ℓ_1 based sparse precision matrix and Gaussian graphical model structure learning methods should be very careful about checking whether the conditions of consistency for precision matrix estimation are likely to be fulfilled in the application area of interest (p. 106).

This finding paralleled Kuismin and Sillanpää (2016), where they similarly noticed inconsistency of the glasso method in that estimation errors did not diminish with increasing sample sizes. Further, Leppä-aho, Pensar, Roos, and Corander (2017) introduced an approximate Bayesian method, and their results showed that glasso was not always consistent with respect to Hamming distance (Norouzi, Fleet, Salakhutdinov, & Blei, 2012). These findings are consistent with results of a less extensive simulation in Epskamp and Fried (2016) and Epskamp (2016) which incidentally also indicated that error did not diminish with larger sample sizes.

Moreover, statistical inference is not straight forward from estimates obtained from ℓ_1-based methods (Hastie, Tibshirani, & Wainwright, 2015, Ch. 6: "Statistical Inference"). That is, just because a variable has been selected, does not allow for claiming the estimate is significantly different from zero, or that a non-selected variable has no-effect. These claims would require formal hypothesis testing (Bayesian or frequentist) (Lockhart, Taylor, Tibshirani, & Tibshirani, 2014; Mohammadi & Wit, 2015; Schäfer & Strimmer, 2005a), which does not equate to selecting variables based on predictive performance or minimizing a particular loss function. For example, selecting a model based on predictive performance can lead to inconsistent model selection (Leng, Lin, & Wahba, 2006). Further, ℓ_1-based methods use automated variable selection, in which valid inference needs to account for model selection bias (Efron, 2014; Lee, Sun, Sun, & Taylor, 2016; Taylor & Tibshirani, 2017), although under certain assumptions "naïve" refitting of the selected variables can lead to valid inferences (S. Zhao, Shojaie, & Witten, 2017). The glasso method faces an additional limitation, because regularization biases the estimates towards zero, which then requires additional steps to obtain nominal frequentist properties (e.g., coverage rates), including debiasing techniques (Javanmard & Montanari, 2015) and non-traditional bootstrapping schemes (Chatterjee & Lahiri, 2011). Together, the central challenge for advancing network methodology in psychology is to not only investigate methods specifically for the most common applications ($p \ll n$), but that also allowing for customary statistical inferences.

In this paper, rather than building upon relatively recently introduced statistical procedures (e.g., ℓ_1-based methods), we propose a statistical technique that directly builds upon work from a century ago (Fisher, 1915, 1924; Yule, 1907), and thus has a closed form solution. We first introduce Gaussian graphical models. We then describe the current default statistical method in psychology, after which we outline our approach for estimating Gaussian graphical models. With a "proof of concept," we demonstrate an important advantage of the proposed method: nominal frequentist properties (e.g., coverage probabilities). We then use simulations to compare the methods with respect to correctly identifying non-zero relationships, in addition to frequentist risk measured with various loss functions. We end with an application to real data, as well as discussing implications and future directions.

Gaussian Graphical Model

Undirected graphical models can refer to covariance selection models, random Markov fields, or network models (as in

psychology). Here we adopted the term Gaussian graphical model (GGM), because it is the most general and provides an informative description of the method. For example, let $\boldsymbol{X}$ be a p-dimensional Gaussian random vector defined as

$$\boldsymbol{X} = \{X_1, ..., X_p\} \sim \mathcal{N}(\mu, \Sigma), \tag{1}$$

where, without loss of generality, we assume all variables have been standardized to have mean μ zero (i.e., $0 = \{\mu_1, ..., \mu_p\}^\top$) and covariance Σ. A GGM is then a probability model that characterizes the conditional dependent structure of $\boldsymbol{X}$ with a graph. This is accomplished by identifying the non-zero elements within the inverse-covariance matrix $\Sigma^{-1} = \boldsymbol{\Theta}$ (i.e., the precision matrix). In the following notation, we denote the graph with $\mathcal{G} = (V, E)$ which consists of nodes $V = \{1, ..., p\}$ as well as the edge set (non-zero connections between nodes) $E \subset V \times V$. The maximum edges possible in $\mathcal{G}$ is $V(V-1)/2$, which correspond to the number of unique off-diagonal elements of Σ. The edge set for $\mathcal{G}$ contains nodes (X_i, X_j) that share a conditional relationship $X_i \not\perp\!\!\!\perp X_j | \boldsymbol{X}_{V \setminus i,j}$. In contrast, conditionally independent nodes $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{V \setminus i,j}$ are not included in E.

The obtained graph $\mathcal{G}$ depends on accurate estimation of the precision matrix $\boldsymbol{\Theta}$. This is straight forward in low-dimensional settings ($p \ll n$), the because maximum likelihood estimator (MLE) provides an adequate estimate (T. Wang et al., 2016). However, in high-dimensional settings ($p \geq n$), the MLE cannot be computed due to singularity: $\det(\Sigma) = 0$. That is, since the determinant equals the product of the eigenvalues (λ)

$$\det(\Sigma) = \prod_{i=1}^{p}, \lambda_i \in \{1, ..., p\}, \tag{2}$$

and the maximum number of non-zero eigen values is $\min(n, p)$ (Kuismin & Sillanpää, 2017), it can be shown that inversion is not possible (Hartlap et al., 2007). This is known as the "large p and small n" problem and remains a central challenge in the field of statistics (Kuismin & Sillanpää, 2016). Although these kinds of data structures are common in fields such as genomics (Y. R. Wang & Huang, 2014) and quantitative finance (Ledoit & Wolf, 2004a, 2004b), they are the exception in psychology. Nonetheless, in psychology, ℓ_1 penalized maximum likelihood has emerged as the default estimation method (Epskamp & Fried, 2016).

ℓ_1 Regularization

In the familiar context of multiple regression, the lasso method uses the ℓ_1 norm to find coefficients that minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where λ is the tuning parameter on the sum of absolute values for the coefficients $|\beta_j|$ (Tibshirani, 1996). Larger values for λ provide more regularization, whereas $\lambda = 0$ results in a non-penalized model. Under the assumption that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, minimizing the ordinary least squares estimates is equivalent to maximizing the likelihood, or in this case the ℓ_1 penalized maximum likelihood. Importantly, optimizing Equation 3 has the ability to reduce coefficients to exactly zero, which allows for variable selection. For this reason ℓ_1-based methods have become popular for both regression and for estimating network models.

Extended to multivariate settings, the penalized likelihood for the precision matrix is defined as

$$l(\boldsymbol{\Theta}) = \log\det\ \boldsymbol{\Theta} - \text{tr}(\boldsymbol{S}\boldsymbol{\Theta}) - \lambda_p \sum_{i \neq j} (|\boldsymbol{\Theta}_{i,j}|) \tag{4}$$

where $\boldsymbol{S}$ is the sample covariance matrix and λ_p a penalty function (Gao, Pu, Wu, & Xu, 2009). The glasso method applies a penalty on the sum of absolute covariance values $\lambda_p(|\boldsymbol{\Theta}_{i,j}|)$ (Friedman et al., 2008). The performance of the glasso method is strongly influenced by the choice of λ, which can be attained in at least four ways: (1) choose λ that minimizes the extended Bayesian information criterion (EBIC; Foygel & Drton, 2010); (2) select λ that minimizes the Rotation Information Criterion (RIC; T. Zhao, Liu, Roeder, Lafferty, & Wasserman, 2012); (3) select λ that maximizes the stability of the solution across subsamples of the data (i.e., Stability Approach to Regularization Selection; StARS) (H. Liu, Roeder, & Wasserman, 2010); and (4) base the selection on k-fold cross-validation (Bien & Tibshirani, 2011).

While a method would ideally be selected with a particular goal in mind, or based on performance in simulations that are representative of the particular field, the default method in psychology is currently EBIC,

$$\text{EBIC} = -2l(\boldsymbol{\Theta}) + E\log(n) + 4\gamma E\log(p), \tag{5}$$

where $l(\boldsymbol{\Theta})$ is defined in Equation 4, E is the size of the edge set (i.e., the number of non-zero elements of $\boldsymbol{\Theta}$), and $\gamma \in [0, 1]$ is the EBIC hyperparameter that puts an extra penalty on the standard Bayesian Information Criterion (BIC) criterion ($\gamma = 0$). The selected network then minimizes EBIC with respect to λ. This is typically accomplished by assessing a large number (e.g., 100) of values of λ and selecting the one for which EBIC is smallest. There is no automatic selection procedure for the EBIC hyperparameter, but 0.5 was recommended in Foygel and Drton (2010) and Epskamp and Fried (2016).

Basic Approach

Our proposed method differs from glasso with several respects. We approach the problem in the simplest terms, in that we are *simply* estimating a (partial) correlation matrix following classic and well known standard methods. Let X_{in} be the observed data for the ith ($i \in 1, ..., p$) variable at the

nth ($n \in 1, ..., N$) observation. We first compute the $p \times p$ covariance matrix with the MLE defined as

$$\Sigma = \frac{1}{N}\sum_{n=1}^{n}(X_n - \bar{X})(X_n - \bar{X})^\top, \quad \bar{X} = \frac{1}{N}\sum_{n=1}^{n} X_n, \tag{6}$$

where the variables are centered to have mean zero. As is evident in Equation 6, this method does not use any form of regularization. After the MLE is computed, it is straight forward to obtain the precision matrix $\Sigma^{-1} = \hat{\boldsymbol{\Theta}}$ which contains the covariances $\hat{\theta}_{ij}$ and variances $\hat{\theta}_{ii}$:

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{\theta}_{ii} & & \\ \vdots & \ddots & \\ \hat{\theta}_{ij} & \cdots & \hat{\theta}_{jj} \end{bmatrix}. \tag{7}$$

The partial correlations can be obtained as

$$\hat{\rho}_{ij} = \frac{-\hat{\theta}_{ij}}{\sqrt{\hat{\theta}_{ii}\hat{\theta}_{jj}}}, \tag{8}$$

which denote the standardized conditional relationships. In contrast to ℓ_1 regularization, where exact zeros are obtained through optimization, this approach requires an explicit decision rule for setting $\hat{\rho}_{ij}$ to zero. Here we first use the Fisher Z-transformation

$$z_{ij} = \frac{1}{2}\log\Big(\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}\Big), \tag{9}$$

which results in an approximate normal distribution defined as

$$z_{ij} \sim \mathcal{N}\left(\frac{1}{2}\log\Big(\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}\Big), \frac{1}{n-3-s}\right). \tag{10}$$

Here s denotes the number of variables controlled for ($p-1$) and $\sqrt{\frac{1}{n-3-s}}$ the standard error. We then define α based on subjective grounds (e.g., the trade off between false positives and negatives), and the corresponding critical value $Z_{\alpha/2}$. In contrast to λ in glasso, α is a calibrated measure with respect to false positives and coverage probabilities–i.e., $100(1-\alpha)$%. The confidence interval for z_{ij} is defined as

$$\begin{aligned} Z_L &= z_{ij} - Z_{\alpha/2}\sqrt{\frac{1}{n-3-s}}, \\ Z_U &= z_{ij} + Z_{\alpha/2}\sqrt{\frac{1}{n-3-s}}, \end{aligned} \tag{11}$$

where Z_L and Z_U denote the lower and upper bounds. To obtain the interval for $\hat{\rho}_{ij}$, a transformation is required:

$$\hat{\rho}_{ij_L} = \frac{\exp(2Z_L)-1}{\exp(2Z_L)+1} \text{ and } \hat{\rho}_{ij_U} = \frac{\exp(2Z_L)-1}{\exp(2Z_U)+1}. \tag{12}$$

From this method, we obtain an edge set E in which the confidence intervals for $\hat{\rho}_{ij}$ exclude zero. If the assumptions of this model are satisfied, the computed intervals will have the nominal coverage probabilities. In the context of Gaussian graphical models, this suggests we can obtain $\approx 100\%$ coverage, or that the false positive rate will be close to zero. For example, specificity (SPC), or true negative rate, is defined as

$$\text{SPC} = \frac{\text{true negative}}{\text{true negative + false positive}}, \tag{13}$$

which should correspond exactly to the coverage rate of $\rho_{ij} = 0$ for a given network. Accordingly, 1 - SPC corresponds to the false positive rate.

Proof of Concept: Coverage Probabilities

In this section, we investigate coverage probabilities of the proposed CI based method. This was done for two reasons: 1) The covariance matrix can be inverted in low-dimensional settings ($p < n$), but there can still be increased estimation errors when p approaches n (Ledoit & Wolf, 2004b); and 2) Alternative approaches, developed for high-dimensional settings in particular, construct an approximate null sampling distribution for the partial correlations, and then use p-values to determine the edge set E (Schäfer & Strimmer, 2005b). To our knowledge, coverage probabilities for partial correlations have not been examined in relatively large p settings.

We simulated data from null networks, in which all partial correlations were set to 0. The corresponding precision matrices $\boldsymbol{\Theta} \sim W_G(df = 20, I_p)$ were then generated from a Wishart distribution with 20 degrees of freedom and scale I_p (Mohammadi & Wit, 2015). The number of variables p was fixed at 20 and the sample sizes varied: $n \in \{25, 50, 150, 250, 500, 1{,}000, \text{and } 10{,}000\}$. The 95 and 99% coverage probabilities were averaged across 1,000 simulation trials. We also plotted results from a representative trial to illustrate coverage for a given network, with the 95% confidence interval, which demonstrates the exact correspondence to specificity.

The results from one simulation trial are plotted in Figure 1. Panel A shows the properties of the computed 95% intervals, in which false positives are denoted in black. Note that the estimated confidence intervals have several desirable characteristics, including being bounded between -1 and 1. When the sample is larger than 25, they are symmetric and become narrower with increasing sample sizes. This stands in contrast to lasso estimation, in which only point estimates are provided by optimizing Equation 4. Further, when bootstrap schemes are used, the sampling distribution can be distorted which is a well-known result of ℓ_1-regularization (Hastie et al., 2015). This point is further discussed and demonstrated in the applied example (Figure 4; Section Application)

The corresponding coverage rates are provide in panel B of figure 1, where the expected level is 95-%. Note that

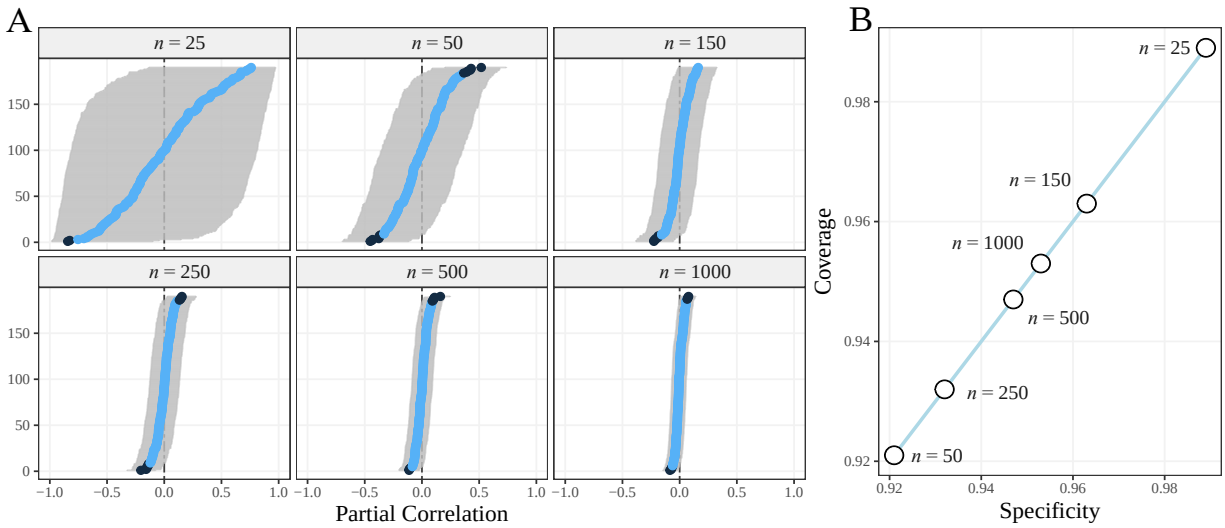

Figure 1. Panel A: Estimated partial correlations and confidence intervals. Black dots denote confidence intervals that exclude zero. Panel B: Specificity and coverage probabilities for the estimated networks in panel A. This demonstrates that, for a given network, specificity and coverage are equivalent.

there is a direct 1:1 correspondence among specificity and coverage illustrated by the diagonal line. This was confirmed with the exact correspondence to specificity (Equation 13), which is a measure of binary classification accuracy that is often used in the GGM literature. It should be noted that coverage was very close to nominal levels, for example, even for one simulation trial it ranged from 93.2% to 96.3% when the samples sizes were larger than 50. With sample sizes of 500 and 1,000, the coverage rate was 94.7% and 95.3%, respectively. Further, as seen in Table 1, long-run coverage probabilities were at the expected levels. This is especially important in applied settings, because it allows for a more principled and familiar rationale for determining the trade-off between false positives and negatives. The current alternative in psychology is to adjust the γ value in EBIC (Equation 5), which paradoxically resulted in diminishing returns with increasing sample sizes (Epskamp & Fried, 2016), in addition to γ not having a straightforward meaning. In contrast, confidence intervals are commonly used, have a straightforward frequentist interpretation (Morey, Hoekstra, Rouder, & Lee, 2015), and allow for defining expected long-run error rates (α).

Simulation Description

In this section, we present numerical experiments to assess performance of the proposed CI method compared to glasso. We specifically focus on model selection consistency in common situations where network models are fit in psychology. We assumed fixed $p = 20$, and increased the sample size $n \in \{50, 150, 250, 500, 1{,}000, 10{,}000 \text{ and } 100{,}000\}$. The largest sample sizes were included to assess consistency of each method. In applied settings, this mimics choosing a psychometric scale (fixed p) and then assessing expected performance by increasing the sample size (n). Additionally, we included a range of sparsity levels, in which the proportion of connected edges varied (0.2, 0.4, 0.6, and 0.8). The edge sets were randomly generated from a Bernoulli distribution, and the corresponding precision matrices $\mathbf{\Theta} \sim W_G(df = 20, A_{p\times p})$ from a Wishart distribution with 20 degrees of freedom and scale A that had 20's along the diagonal and 0's as the off-diagonal elements. This choice of df ensured the partial correlations were within a reasonable range ($\rho_{ij} \approx \pm\, 0.40$), in addition to being approximately normally distributed with mean zero. This scale ($A_{p\times p}$) differed from Mohammadi and Wit (2015), who used an identity matrix I_p, but was selected to provide the most favorable conditions for the glasso method, which we noted had worse performance (specifically for the risk of $\hat{\mathbf{\Theta}}$) when the diagonal of the true precision contained too large or small values.

We used the package *qgraph* to fit the glasso models (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012). Here we assumed γ values of 0 and 0.5. The latter is the default setting in *qgraph*. For the largest sample sizes, we

Table 1
Average coverage probabilities for the partial correlation matrices. The parentheses include the standard deviations.

	Sample size (n)					
	25	50	150	250	500	1000
95-% CI	0.970	0.953	0.952	0.951	0.951	0.950
	(0.06)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
99-% CI	0.995	0.990	0.990	0.990	0.990	0.990
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

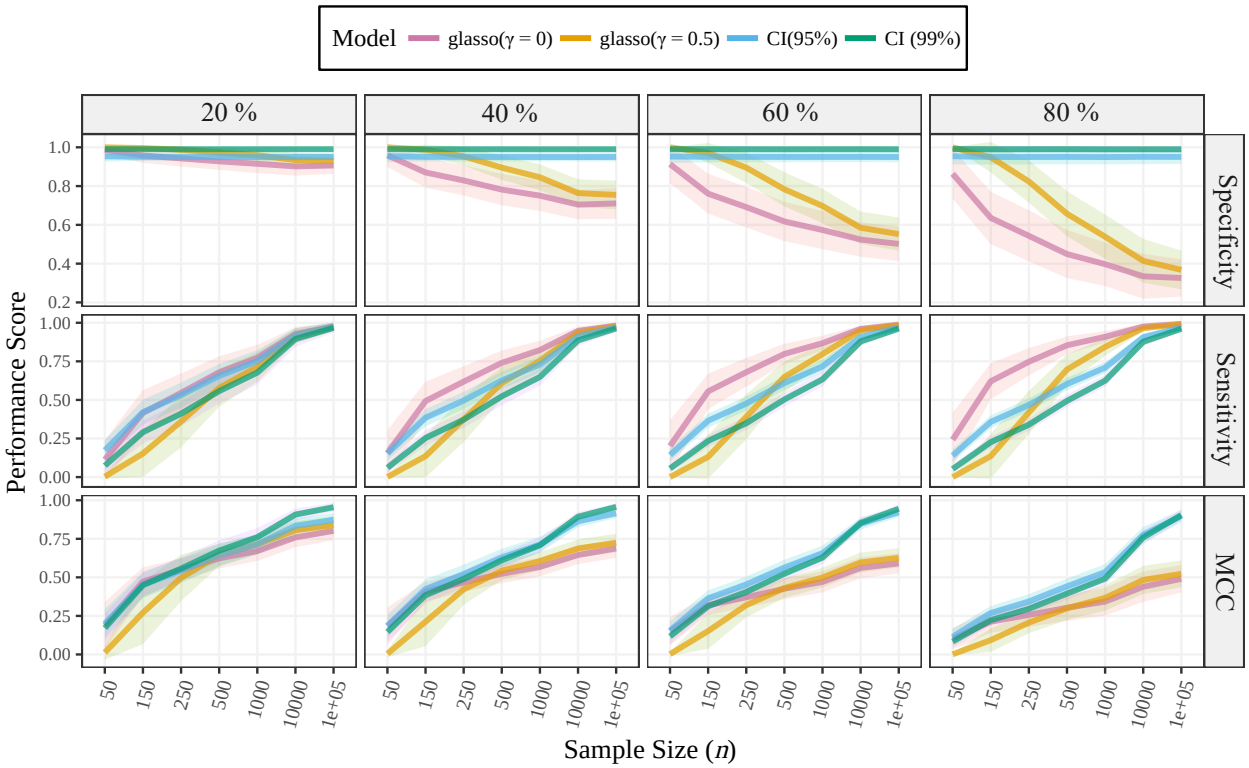

Figure 2. Edge set identification scores (1 - Specificity is the false positive rate).The probability of connection is denoted with 20 %, 40 %, 60 %, and 80 %. MCC = Mathews correlation coefficient.

observed warnings that the lowest λ values was chosen (Section ℓ_1 Regularization). We followed the package recommendation, and allowed this settings to be changed during the simulation. For our proposed CI method, we used two confidence levels of 95-% and 99-%. These models were fit with a custom function that is provided in the Appendix. The performance measures were averaged across 1,000 simulation trials. All computer code is publicly available on the Open Science Framework (link).

Edge Set Identification

We assessed three measures for identifying non-zero partial correlations. The first was specificity, which was previously defined in Equation 13. The next measure is sensitivity (SN), or the true positive rate, and is defined as

$$\text{SN} = \frac{\text{true positives}}{\text{true positive + false negatives}}. \tag{14}$$

We also wanted to include a measure that considers all aspects of binary classification (i.e., false positives (FP) and negatives (FN), as well as true positives (TP) and negatives (TP)). To our knowledge, the Matthews correlation coefficient (MCC) is the only measure that meets this criteria. MCC is defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{15}$$

and ranges between -1 and 1 (Powers, 2011). A correlation of 1 is perfect correspondence between the actual and estimated edges. Its value is equivalent to the phi coefficient that assesses the association between two binary variables, but for the special case of binary classification accuracy.

The results are presented in Figure 2. We first discuss specificity (1 - specificity = false positive rate). All methods had similar performance when 20% of the nodes shared a connection. However, while not a large difference, specificity decreased slightly when the sample size (n) grew larger for both glasso models ($\gamma = 0$ and 0.5). This result was especially pronounced for denser networks. For example, with 60% connectivity, the specificity for glasso was 100% ($n = 25$) but was below 80% with a sample size of 500 and approached 50% with $n = 10,00$. In contrast,the proposed CI method (Section Basic Approach) performed uniformly across conditions. Indeed, these results confirm Figure 1 and table 1 where confidence levels corresponded exactly to specificity.

Importantly, the high specificity of the proposed CI method did not result in substantially lower sensitivity than the glasso models. For glasso with $\gamma = 0.5$ in particular, sensitivity was comparable to the CI method, but the false

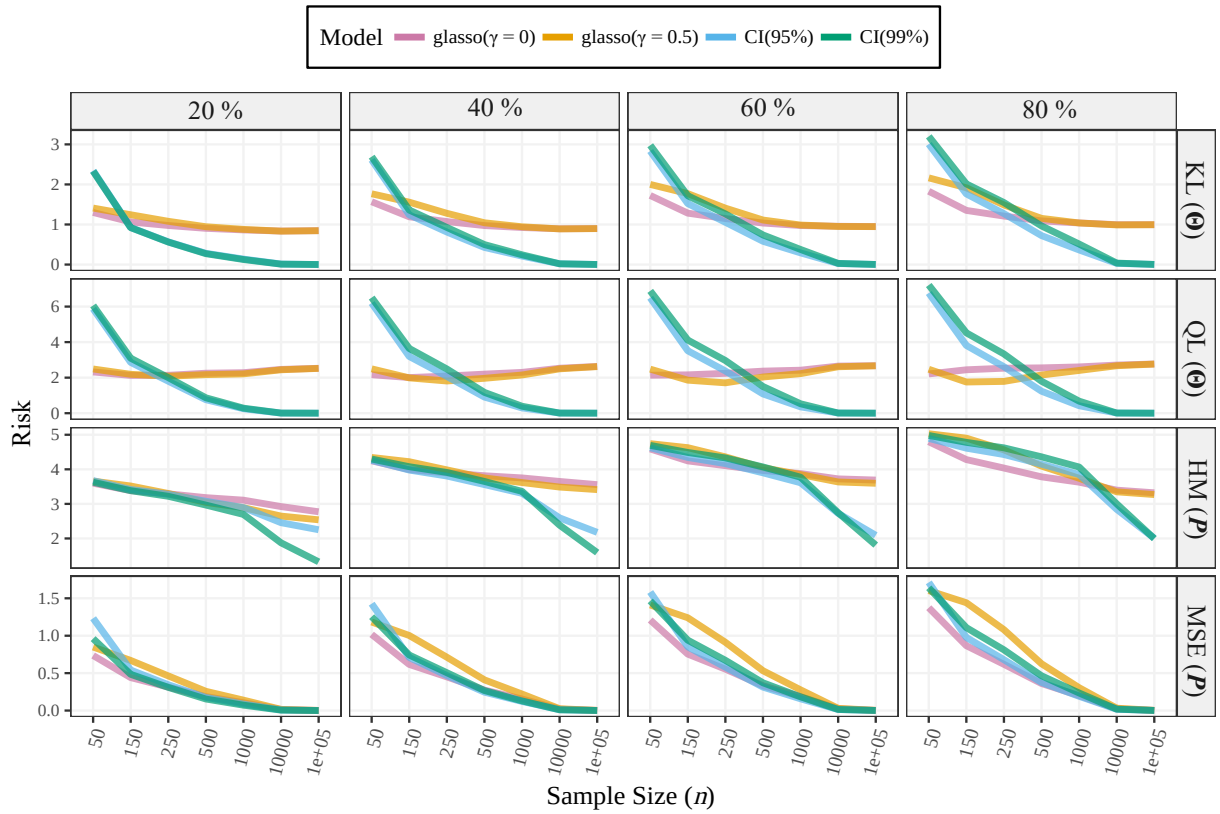

Figure 3. Risk of the estimated precision $\boldsymbol{\Theta}$ *and partial correlation* $\boldsymbol{P}$ *matrices.* γ *denotes the parameter in EBIC (Equation 5). The probability of connection is denoted with 20%, 40%, 60%, and 80%. KL = Kullback-Leibler divergence; QL = quadratic loss; HM = Hamming distance; MSE = mean squared error. The results are presented on the log(x + 1) scale.*

positive rate was much higher and became increasingly so with larger sample sizes. These results parallel Heinävaara et al. (2016) and Kuismin and Sillanpää (2016), where it was noted that glasso behaves inconsistently as n increases. That is, when sample sizes increase, a reasonable expectation is that the estimated model will become more reliable, but this is not always the case for the glasso method. Further, in the present simulation conditions the proposed CI method turned out to be a consistent estimator for the purpose of model selection. That is, specificity can be set to $\approx$ 99% (i.e., the confidence level), and increasing n ultimately results in selecting the *true* model with a probability approaching 100%.

In terms of the Mathews correlation coefficient (MCC; Equation 15), which provides a correlation for binary variables, all methods performed similarly for a network with 20% connectivity, which parallels the results for specificity and sensitivity. However, the CI based methods often outperformed both glasso methods in the other conditions, although the MCC correlation increased with larger n in all cases. For example, the methods were similar for the smaller sample sizes, but the proposed CI methods resulted in larger correlations with increasing sample sizes. For the largest sample size (n = 10,000) and 60% connectivity, the CI methods had an almost perfect MCC score, whereas the glasso methods had a score of approximately 0.50.

Loss Functions

To further assess the quality of the estimation methods, we compared the glasso and CI methods in terms of risk. Risk of the estimated precision matrices was assessed with two loss functions, each of which are commonly used in the Gaussian graphical model literature. The first is Kullback–Leibler divergence, or entropy loss, defined as

$$\text{KL}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}) = \text{tr}(\boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}}) - \log(|\boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}}|) - p, \tag{16}$$

where $\log(|\boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}}|)$ denotes the log determinant. This provides a measure of information loss between the estimated and *true* model. We also assessed Quadratic loss (QL) that follows

$$\text{QL}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}) = \text{tr}(\boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}} - I_p)^2, \tag{17}$$

where I_p is an identity matrix. Moreover, we assessed the risk of the estimated partial correlation matrices $\boldsymbol{P}$. As a measure of discrepancy among the true and estimated model we also

computed the Hamming distance (Heinävaara et al., 2016), which provides a measure of discrepancy between binary strings. Here, non-zero partial correlations were denoted with 1, whereas the partial correlations that were set to zero were denoted with a 0. Hamming distance (HM) addresses the discrepancy among the true and estimated model. For example,

$$\begin{aligned} \text{True: } & \boxed{00}10\boxed{1}0100, \\ \text{Estimated: } & \boxed{11}10\boxed{0}0100, \end{aligned} \tag{18}$$

results in a Hamming distance of three. The second measure of risk was squared error

$$\mathrm{R}(\boldsymbol{P}, \hat{\boldsymbol{P}}) = \sum_{i \neq j} (\hat{\rho}_{ij} - \rho_{ij})^2, \tag{19}$$

where this value was averaged across the simulations trials. This resulted in mean squared error MSE($\boldsymbol{P}$, $\hat{\boldsymbol{P}}$). For all loss functions, values closer to zero indicate less error from the actual precision or partial correlation matrix.

The results are presented in Figure 3. Before discussing these results in detail, it should be noted that there were some difficulties computing KL-divergence. This occurred with the smallest samples size (n = 50) in particular, and was due to the risk being assessed from the sparsified precision matrix. We revisit this issue in the discussion (Section Limitations). In terms of KL-divergence, both glasso estimates (γ = 0 and 0.5) were inconsistent, in that risk appeared to plateau and did not reduce further with larger sample sizes. Importantly, glasso did have superior performance with the smallest sample sizes (p/n = 0.40), while the CI models had consistent performance. For example, as n increased, the risk consistently diminished to almost zero for the CI based methods. While the results for quadratic loss (QL) showed a similar pattern, the risk actually *increased* for both glasso methods as the sample sizes increased. In contrast, the proposed CI method was again consistent, although glasso did have superior performance for the smaller sample sizes ($n < 250$).

We now describe the results for $\boldsymbol{P}$. For Hamming distance, error reduced for all models. While the models had similar trajectories, it should be noted that only CI based methods approached a Hamming distance of 0. For 60% connectivity, as an example, the glasso methods had risk of almost 50, whereas the proposed method had close to zero errors. Importantly, mean squared error (MSE) was similar for each method. However, for sample sizes less than 1,000, glasso (γ = 0.5) had the highest MSE. This is notable, because this γ value is the default in the R package *qgraph*.

Application

In this section, we estimate the network structure of post-traumatic stress disorder symptoms (Epskamp & Fried, 2016). Our interest is not in a substantive question, but to compare the methods in two quantitative aspects: (1) agreement (or disagreement) between methods, and in particular the degree of estimated sparsity; and (2) to highlight post-selection estimates of the partial correlations, for example, bootstrapping the glasso models compared to the CI based approach. The data consists of 20 variables (p) and 221 observations (n) measured on the likert scale (0 – 4). We thus assumed normality for the CI based methods, while the glasso methods estimated polychoric partial correlations, which is the default approach in the package *qgraph*.[1]

The results are presented in Figure 4. We first discuss the estimated network structures in panel A. There are substantial differences between the methods, in that the glasso estimated dense networks where almost half of the possible edges were connected. In contrast, the CI methods had connectivity of 36% (CI 95-%) and 11% (CI 99-%), respectively. In addition to the simulations presented here (Figure 2), the glasso estimate (γ = 0.5) of these exact data was used to provide the data generating matrix in Epskamp and Fried (2016). The limited simulation provided in Epskamp and Fried (2016) showed that glasso was similarly inconsistent, which parallels the present simulation results, and that specificity was never higher than 75%. This suggests that the estimated network in this example has a false positive rate (1 - specificity) of close to 25% (i.e., 1 out of 4). In contrast, the proposed method not only had the highest specificity (and thus the lowest false positive rate), but similar sensitivity to the glasso methods in this simulation, which together suggests a more accurate estimate of the network.

We now focus on post-selection assessment of the partial correlations for the glasso method (Figure 4; panel B). That is, after glasso has selected a model, common practice in psychology is to use a bootstrapping procedure to approximate the sampling distributions. We thus implemented the default approach in the R package *bootnet* (Epskamp, Borsboom, & Fried, 2018). However, to be clear, the naïve use of bootstrapping does not necessarily allow for valid inferences such as null hypothesis testing with well-defined error rates (α). This is evident in panel B of Figure 4, where it can be seen that the bootstrapped estimates (summarized with the mean and 95-% quantile intervals) are heavily skewed for the default glasso method (γ = 0.5). In the context of Gaussian graphical models in particular (Janková & van de Geer, 2017), statistical inference is an emerging area of research in the field of statistics that often require debiasing the regularized estimates to compute confidence intervals (Janková & van de Geer, 2015; Ren, Sun, Zhang, & Zhou, 2015) and p-values (W. Liu, 2013; Z. Wang, 2016). There is a recent R package– Statistical Inference of Large-Scale Gaussian Graphical Model

[1] We confirmed that the CI based methods generally had nominal coverage for ordinal data that was generated with the *bootnet* package.

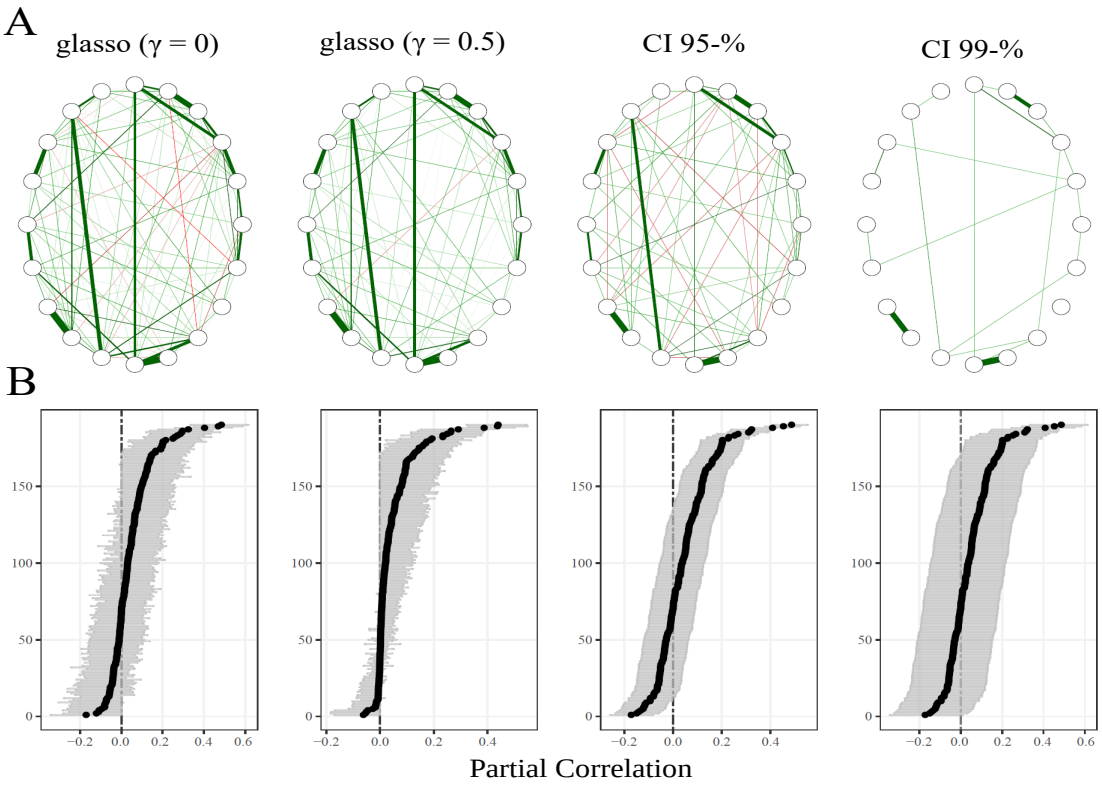

Figure 4. Panel A: Estimated networks for symptoms of post-traumatic stress disorder. γ denotes the parameter in EBIC (Equation 5). Panel B: A comparison between estimates, including confidence intervals, for the estimated networks. The glasso confidence intervals were constructed with a bootstrap procedure from the package bootnet.

in Gene Networks (*SILGGM*)–that provides many options for network model inference (Rong, Ren, & Chen, 2017). However, it should be noted that the methods were optimized in high-dimensional settings ($n < p$), so performance confirmation would be needed in low-dimensional settings. In contrast, because typical psychological networks are fit in low-dimensional settings, the proposed CI method already allows for calibrated confidence intervals (and p-values; Figure 1).

Moreover, we see that the CI based methods, described in the present paper have symmetric intervals that readily allow for demonstrating nominal frequentist calibration (Figure 1 and Table 1). While there is still the issue of multiple comparisons, one could argue that 99% intervals mitigate these multiplicities, without further reducing sensitivity and because increasing the confidence levels results in trivial changes in the width of the intervals. Further, assuming the null is true for each partial correlation, coverage (or non-coverage of 0) and thus specificity can be inferred due to the large number of constructed intervals (Figure 1). This again stands in contrast to glasso with EBIC selection of the tuning parameter (Equation 4), where the meaning of γ is unclear, in addition to the assumed γ (0 and 0.5) values estimating very similar networks.

Discussion

In this paper, we have described the current default approach for estimating psychological networks, with a particular focus on the disconnect between the fields where glasso was developed ($n \ll p$) and the most common psychological applications ($n \gg p$). We then described a method based on maximum likelihood and Fisher Z-transformed partial correlations. With confidence intervals as the decision rule for determining non-zero relationships, we then demonstrated superior performance compared to the glasso method in almost all instances (Figure 2). In particular, we showed the exact correspondence between the confidence level and specificity, which is due to 1 - specificity denoting the false positive rate (e.g., α; Figure 1). As indicated by Figure 3, it is also clear that the glasso method does not reduce risk of the estimated precision matrices, relative to the non-regularized method based on maximum likelihood. Indeed, the glasso methods actually showed increased estimation errors for quadratic loss when the sample sizes became larger. Most importantly, we explicitly evaluated model selection consistency of the glasso method. Here it was shown that glasso

is not a consistent estimator for the purpose of model selection, in low-dimensional settings, whereas the proposed method converged on the *true* model with a probability that approached 100% (Figure 2).

Although our focus here is statistical methodology, and not on the use or corresponding inferences in practice, these results can be used to inform the current discussion surrounding the replicability of psychological networks (Forbes, Wright, Markon, & Krueger, 2017). There are several less extensive simulations that have demonstrated that glasso is not consistent for the purpose of model selection in psychological settings. In fact, we have not seen one instance in which glasso converged upon the *true* model. For example, in Epskamp and Fried (2016) and Epskamp (2016), it was shown that specificity either reduced slightly or remained constant at around 75 − 80% as n increased. That is, the false positive rate (1 - specificity) of glasso is regularly around 20 − 25%. Further, while Epskamp et al. (2017) cautioned that assuming sparsity will result in false negatives if the *true* network is dense, our results suggest that levels of sparsity not typically seen in psychological applications (< 20% connectivity; Figure 2) are necessary for consistent model selection (although specificity declined slightly for the largest sample sizes). In the context of replication, high false positive rates (> 20%) obscure the ability to consistently replicate network structures. Although the glasso method appears to estimate similar networks across datasets (Fried et al., 2018), for example, it is not entirely clear what is being replicated for a method whose performance is consistently inconsistent (Epskamp, 2016; Epskamp & Fried, 2016; Heinävaara et al., 2016; Kuismin & Sillanpää, 2016; Leppä-aho et al., 2017).

These results may be surprising to some, because the glasso method has emerged as the default approach for network estimation in psychology. However, while the original glasso paper is highly cited (Friedman et al., 2008), it should be noted that the performance of the method for edge identification was not assessed. Similarly, in Foygel and Drton's (2010) work that introduced EBIC for tuning parameter (λ) selection, no comparison to other methods was made. However, there are numerous papers that have demonstrated superior performance than glasso with EBIC (Kuismin & Sillanpää, 2017, see here for a review of different methods). For example, Leppä-aho et al. (2017) introduced an approximate Bayesian method, using a marginal pseudo-likelihood approach, that showed glasso was not always consistent with respect to Hamming distance (Norouzi et al., 2012), whereas the LASSO regression approach was consistent (Meinshausen & Bühlmann, 2006). This finding parallels Kuismin and Sillanpää's (2016), where the unusual behavior of glasso was explicitly noted:

> We are surprised by the moderate performance of the graphical lasso in this simulation setting. Even when the sample size increases, the risk measures do not diminish, and that is quite unexpected. This is most certainly due the EBIC used to choose the regularization parameter ρ [λ] (p.12).

Again, these methods were developed for high-dimensional settings, and thus the focus was not on low-dimensional settings where classic methods are performing well. In fact, most common statistical methods (e.g., maximum likelihood) are known to have optimal performance in situations common to psychology. In this light, it is clear that the results presented in the current paper are not too surprising if viewed from the position of going "back to the basics." That is, in most psychological applications, partial correlation networks are most *simply* estimating correlation matrices in settings that do not pose challenges for statistical approaches developed over a century ago. Of course, while using a Fisher Z-transformation does not have the appeal of novelty as glasso, regularization, or EBIC, it is also clear that going "back to the basics" provides consistent model selection in the most common situations were psychological networks are estimated.

Limitations

There are several limitations of this work. First, predictive accuracy is one possible advantage of ℓ_1 regularization, but we did not consider this here. However, it should be noted that ℓ_1-based methods do not always have improved predictive accuracy. For example, according to Friedman et al. (2008, the original glasso paper) "...cross-validation curves indicate that the unregularized model is the best, [which is] not surprising given the large number of observations and relatively small number of parameters" (p. 9). Nonetheless, alternative methods based on non-regularized regression models could be used to select variables with the Bayesian information criterion, which is known to be consistent for model selection (Casella et al., 2009, $p \ll n$) and can be justified in terms of predictive accuracy (Shao, 1997, leave–v–out). Second, we only considered networks with a random structure. Future work would have to evaluate whether these findings generalize to various network structures, which seems reasonable since the proposed method is based on maximum likelihood (Equation 6). Third, while the confidence intervals are calibrated (Figure 1), it is not clear how to statistically compare partial correlations to each other. Although the bootstrap approach is recommended in Epskamp and Fried (2016), we were unable to locate any proofs in the statistics literature that this procedure generally allows for valid inferences. In fact, according to Bühlmann, Kalisch, and Meier (2014):

> ...[W]e typically use sparse estimators for high-dimensional data analysis, for example the Lasso... The (limiting) distribution of such a

> sparse estimator is non-Gaussian with point mass at zero, and this is the reason why standard bootstrap or subsampling techniques do not provide valid confidence regions or p-values. Thus, we have to use other approaches to quantify uncertainty (p. 7–8).

Rather than attempting to overcome the biased estimates of ℓ_1 regularization, a non-regularized bootstrap could be applied directly on the maximum likelihood estimator (Equation 6), from which differences as well as equivalence can be tested (Lakens, 2017). Of course, this would first require demonstrating that the constructed intervals and/or p-values are properly calibrated. Fourth, we only evaluated simulation conditions with p fixed to 20. While this is a reasonable choice based on the psychological literature (Epskamp & Fried, 2016), it should be noted that estimation errors of the MLE arise with larger p/n ratios. However, for the purpose of edge set identification, the CI based methods outperformed the glasso ($\gamma = 0.5$; the default in *qgraph*) with the highest ratio evaluated ($p/n = 0.40$). Fifth, the proposed method had difficulties computing KL-divergence. In the context of determining non-zero partial correlations, in which sparsity is induced after Equation 8, this is not problematic. This issue arose because KL-divergence was assessed with covariances forced to zero, which we viewed as a fairer comparison to the glasso method (that also has covariances set to zero) and allowed for assessing risk for each confidence level (using the non-sparsified precision matrix would have provided the same estimate for each decision rule). Importantly, in all instances the estimated precision matrices were positive definite.

To be clear, while not necessary a limitation of this work, it should be noted that we used the default settings in the package *qgraph*. This allowed for making our findings especially relevant for psychology, but does limit the generalizability of our results. For example, there are alternative default settings in other R packages (T. Zhao et al., 2012, e.g., *huge*), where EBIC is not the default method for selecting λ. We did explore many of the settings for the glasso method. For example, in addition to different methods for selecting λ, the range of λ's can change the results in meaningful ways. If the true model is known, it is possible to adapt a number of parameter settings to improve performance in glasso. However, we view this as an additional benefit of the proposed method, because performance only depends on pre-specifying the confidence level which has a straight forward meaning in practice.

Conclusion

To be clear, Gaussian graphical models are useful tools in that they can provide important insights into psychological phenomena. These insights are not possible with more traditional statistical techniques such as structural equation models. An important future direction is therefore to address the issues that we raised, in addition to further characterizing non-regularized methods, which together will provide a deeper understanding of this relatively novel approach for conceptualizing a correlation matrix. However, in regards to the current default approach in psychology, we believe the statistical foundations of partial correlation network methodology requires rethinking.

Appendix A
R-code

```
mle_CI <- function(X, alpha){
  X <- as.matrix(X)
  # X: data frame

  if (!require("qgraph")) install.packages("
      qgraph")
  if (!require("Matrix")) install.packages("
      Matrix")
  # number of observations (rows)
  n <- nrow(X)
  # number of variables (columns)
  p <- ncol(X)
  ## compute maximum likelihood estimator
  ## for covariance matrix
  mle_cov <- crossprod(scale(X, scale = F))
      / n

  ## compute maximum likelihood estimator of
       precision matrix
  ## (inverse covariance matrix)
  mle_inv <- solve(mle_cov)

  ## standardize and revese sign = partial
      correaltions
  par_cors  <-  as.matrix(qgraph::wi2net(mle
      _inv))



  mle_parcors <- mle_ci_helper(alpha = alpha
      , par_cors = par_cors, n = n, s = p -
      1)

  mle_inv <- mle_parcors$sig_mat * mle_inv
  list(mle_parcors = mle_parcors, mle_inv =
      mle_inv)

}


mle_ci_helper <- function(alpha, par_cors, s
    , n) {

  # n: sample size
  # s: p - 1 (controlled for)
  # alpha: confidence level
```

```
  # par_cors: partial correlations
  mat <- matrix(0,nrow =  s + 1, ncol = s +
      1)
  CI_ls <- list()
  par_cor <- par_cors[upper.tri(par_cors)]
  cov <- list()
  for(i in 1:length(par_cor)){
    # critical value
    z_crit <- qnorm(1 - alpha/2)
    # standard error
    se <- sqrt(1/((n - s - 3)))
    # z transformation
    z <- log((1 + par_cor[i])/(1 - par_cor[i
        ]))/2
    # z lower bound
    Z_L <- z - z_crit * se
    # Z upper bound

    Z_U <- z + z_crit * se
    rho_L <- (exp(2*Z_L) - 1)/(exp(2*Z_L) +
        1)
    rho_U <- (exp(2*Z_U) - 1)/(exp(2*Z_U) +
        1)
    CI <- c(rho_L, rho_U)
    CI_ls[[i]] <- CI
    cov[[i]] <- ifelse(CI[1] < 0 & CI[2] >
        0, 0, 1)
  }

  ci_dat <- do.call(rbind.data.frame, CI_ls)
  colnames(ci_dat) <- c("low", "up")
  ci_dat$pcor <- unlist(par_cor)
  diag(mat) <- 1
  mat[upper.tri(mat)] <- unlist(cov)
  mat <- as.matrix(Matrix::forceSymmetric(
      mat))
  list(sig_mat = mat, par_cors = par_cors,
      par_sig = mat * par_cors,
        cis = ci_dat, cov_prob = unlist(cov))
}
Assume X is a data matrix:
# 95 % CI
est_mle_95 <- mle_CI(X, alpha = 1 - 0.95)
# sparsified partial correlation matrix
est_mle_95$mle_parcors$par_sig

# 99 % CI
est_mle_99 <- mle_CI(X, alpha = 1 - 0.99)
# sparsified partial correlation matrix
est_mle_99$mle_parcors$par_sig
```

Appendix B
Open Science Framework

Here: link to project

References

Bien, J., & Tibshirani, R. J. (2011, 12). Sparse estimation of a covariance matrix. *Biometrika*, *98*(4), 807–820. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23049130http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3413177 doi: 10.1093/biomet/asr054

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev-clinpsy-050212-185608 doi: 10.1146/annurev-clinpsy-050212-185608

Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-Dimensional Statistics with a View Toward Applications in Biology. *Annual Review of Statistics and Its Application*, *1*(1), 255–278. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev-statistics-022513-115545 doi: 10.1146/annurev-statistics-022513-115545

Casella, G., Girón, F. J., Martinez, M. L., & Moreno, E. (2009). Consistency of bayesian procedures for variable selection. *Annals of Statistics*, *37*(3), 1207–1228. doi: 10.1214/08-AOS606

Chatterjee, A., & Lahiri, S. N. (2011). *Bootstrapping Lasso Estimators* (Vol. 106). Taylor & Francis, Ltd.American Statistical Association. Retrieved from http://www.jstor.org/stable/41416396 doi: 10.2307/41416396

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29. Retrieved from http://dx.doi.org/10.1016/j.jrp.2014.07.003 doi: 10.1016/j.jrp.2014.07.003

Dalalyan, A. S., Hebiri, M., & Lederer, J. (2017). On the prediction performance of the Lasso. *Bernoulli*, *23*(1), 552–581. Retrieved from http://projecteuclid.org/euclid.bj/1475001366 doi: 10.3150/15-BEJ756

Dempster, A. (1972, 3). Covariance Selection. *Biometrics*, *28*(1), 157–175. Retrieved from http://www.jstor.org/stable/2528966?origin=crossref doi: 10.2307/2528966

Efron, B. (2014, 7). Estimation and Accuracy After Model Selection. *Journal of the American Statistical Association*, *109*(507), 991–1007. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.823775 doi: 10.1080/01621459.2013.823775

Epskamp, S. (2016). Regularized Gaussian Psychological Networks: Brief Report on the Performance of Extended BIC Model Selection. *arXiv*, 1–6. Retrieved from http://arxiv.org/abs/1606.05771

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. doi: 10.3758/s13428-017-0862-1

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, *48*(4). Retrieved from http://www.jstatsoft.org/v48/i04/ doi: 10.18637/jss.v048.i04

Epskamp, S., & Fried, E. I. (2016). A Tutorial on Regularized Partial Correlation Networks. *arXiv*. Retrieved from http://arxiv.org/abs/1607.01367 doi: 10.1103/PhysRevB.69.161303

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PLoS ONE*, *12*(6), 1–13. doi: 10.1371/journal.pone.0179891

Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econometrics Journal*, *19*(1), C1-C32. doi: 10.1111/ectj.12061

Fisher, R. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507–521.

Fisher, R. (1924). *The distribution of the partial correlation coefficient.* (Vol. 3).

Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*(7), 969–988. doi: 10.1037/abn0000276

Foygel, R., & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. *Advances in Neural Information Processing Systems*, 604–612. Retrieved from http://arxiv.org/abs/1011.6640%5Cnhttp://papers.nips.cc/paper/4087-extended-bayesian-information-criteria-for-gaussian-graphical-models

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., … Karstoft, K.-I. (2018). Replicability and Generalizability of Posttraumatic Stress Disorder (PTSD) Networks: A Cross-Cultural Multisite Study of PTSD Symptoms in Four Trauma Patient Samples. *Clinical Psychological Science*, 216770261774509. Retrieved from http://journals.sagepub.com/doi/10.1177/2167702617745092 doi: 10.1177/2167702617745092

Friedman, J., Hastie, T., & Tibshirani, R. (2008, 7). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. Retrieved from https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxm045 doi: 10.1093/biostatistics/kxm045

Gao, X., Pu, D. Q., Wu, Y., & Xu, H. (2009). Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. *arXiv*. Retrieved from http://arxiv.org/abs/0909.0934 doi: 10.5705/ss.2009.210

Ha, M. J., & Sun, W. (2014). Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. *Biometrics*, *70*(3), 765–773. doi: 10.1111/biom.12186

Hartlap, J., Simon, P., & Schneider, P. (2007). Unbiased estimation of the inverse covariance matrix. *a&a*, *464*, 399–404. Retrieved from https://www.aanda.org/articles/aa/pdf/2007/10/aa6170-06.pdf doi: 10.1051/0004-6361:20066170

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations.* Boca Raton: CRC Press. doi: 10.1201/b18401-1

Heinävaara, O., Leppä-aho, J., Corander, J., & Honkela, A. (2016). On the inconsistency of L1-penalised sparse precision matrix estimation. *BMC Bioinformatics*, *17*(Suppl 16). doi: 10.1186/s12859-016-1309-x

Janková, J., & van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, *9*, 1205–1229. doi: 10.1214/15-EJS1031

Janková, J., & van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, *26*(1), 143–162. doi: 10.1007/s11749-016-0503-5

Javanmard, A., & Montanari, A. (2015). De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. *arXiv*. Retrieved from http://arxiv.org/abs/1508.02757

Kuismin, M., & Sillanpää, M. (2017). Estimation of covariance and precision matrix, network structure, and a view toward systems biology. *Wiley Interdisciplinary Reviews: Computational Statistics*, *9*(6), 1–13. doi: 10.1002/wics.1415

Kuismin, M., & Sillanpää, M. J. (2016). Use of Wishart prior and simple extensions for sparse precision matrix estimation. *PLoS ONE*, *11*(2), e0148171. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4734711 doi: 10.1371/journal.pone.0148171

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/1948550617697177https://osf.io/preprints/psyarxiv/97gpc/ doi: 10.1177/1948550617697177

Ledoit, O., & Wolf, M. (2004a). Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management*, *30*(4), 110–119. Retrieved from http://www.iijournals.com/doi/abs/10.3905/jpm.2004.110 doi: 10.3905/jpm.2004.110

Ledoit, O., & Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. doi: 10.1016/S0047-259X(03)00096-4

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, *44*(3), 907–927. doi: 10.1214/15-AOS1371

Leng, C., Lin, Y., & Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, *16*(1091), 1273–1284. Retrieved from http://www.stat.wisc.edu/~wahba/stat840/pdf1/leng.lin.wahba.modsel.06.pdf

Leppä-aho, J., Pensar, J., Roos, T., & Corander, J. (2017). Learning Gaussian graphical models with fractional marginal pseudo-likelihood. *International Journal of Approximate Reasoning*, *83*, 21–42. Retrieved from http://dx.doi.org/10.1016/j.ijar.2017.01.001 doi: 10.1016/j.ijar.2017.01.001

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *arXiv*, 1–14. Retrieved from http://arxiv.org/abs/1006.3316 doi: papers3://publication/uuid/F1CE0C72-5199

-4FC6-829C-B76A36C5ED28

Liu, W. (2013). Gaussian Graphical Model Estimation with False Discovery Rate Control. *The Annals of Statistics*, *41*(6), 2948–2978. Retrieved from http://arxiv.org/abs/1306.0976 doi: 10.1214/13-AOS1169

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, *42*(2), 413–468. doi: 10.1214/13-AOS1175

Maccallum, R. C., & Austin, J. T. (2000). APPLICATIONS OF STRUCTURAL EQUATION MODELING IN PSYCHOLOGICAL RESEARCH. *Annual review of psychology*, *1*, 201–226. doi: 10.1146/annurev.psych.51.1.201

McNally, R. J., Robinaugh, D. J., Wu, G. W. Y., Wang, L., Deserno, M. K., & Borsboom, D. (2015, 11). Mental Disorders as Causal Systems. *Clinical Psychological Science*, *3*(6), 836–849. Retrieved from http://journals.sagepub.com/doi/10.1177/2167702614553230 doi: 10.1177/2167702614553230

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, *34*(3), 1436–1462. doi: 10.1214/009053606000000281

Mohammadi, A., & Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, *10*(1), 109–138. doi: 10.1214/14-BA889

Morey, R. D., Hoekstra, R., Rouder, J. N., & Lee, M. D. (2015, 10). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & ...*, *23*(1), 103–23. Retrieved from http://dx.doi.org/10.3758/s13423-015-0947-8VN-readcube.com doi: 10.3758/s13423-015-0947-8

Norouzi, M., Fleet, D. J. D. D. J., Salakhutdinov, R., & Blei, D. M. (2012). Hamming distance metric learning. *Nips*, 1–9. Retrieved from https://papers.nips.cc/paper/4808-hamming-distance-metric-learning.pdf

Peng, J., Wang, P., Zhou, N., & Zhu, J. (2009, 6). Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association*, *104*(486), 735–746. Retrieved from http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0126 doi: 10.1198/jasa.2009.0126

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. Retrieved from http://www.bioinfopublication.org/files/articles/2_1_1_JMLT.pdf doi: 10.1.1.214.9232

Ren, Z., Sun, T., Zhang, C.-H., & Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, *43*(3), 991–1026. Retrieved from http://projecteuclid.org/euclid.aos/1431695636 doi: 10.1214/14-AOS1286

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence*, *161*, 230–237. Retrieved from http://dx.doi.org/10.1016/j.drugalcdep.2016.02.005 doi: 10.1016/j.drugalcdep.2016.02.005

Rong, Z., Ren, Z., & Chen, W. (2017). *SILGGM: Statistical Inference of Large-Scale Gaussian Graphical Model in Gene Networks*. Retrieved from https://cran.r-project.org/package=SILGGM

Schäfer, J., & Strimmer, K. (2005a, 3). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, *21*(6), 754–764. Retrieved from https://www.aanda.org/articles/aa/pdf/2007/10/aa6170-06.pdf doi: 10.1093/bioinformatics/bti062

Schäfer, J., & Strimmer, K. (2005b). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). Retrieved from https://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1175/sagmb.2005.4.1.1175.xml doi: 10.2202/1544-6115.1175

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*, 221–264. Retrieved from http://www3.stat.sinica.edu.tw/statistica/password.asp?vol=7&num=2&art=1

Taylor, J., & Tibshirani, R. (2017). Post-selection inference for ℓ1-penalized likelihood models. *Canadian Journal of Statistics*(2015), 1–26. doi: 10.1002/cjs.11313

Tibshirani, R. (1996). *Regression Selection and Shrinkage via the Lasso* (Vol. 58) (No. 1). Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574 doi: 10.2307/2346178

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*, 1–10. doi: 10.1038/srep05918

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., ... Chen, W. (2016). FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLoS Computational Biology*, *12*(2), 1–16. doi: 10.1371/journal.pcbi.1004755

Wang, Y. R., & Huang, H. (2014, 12). Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, *362*, 53–61. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022519314001969http://linkinghub.elsevier.com/retrieve/pii/S0022519314001969 doi: 10.1016/j.jtbi.2014.03.040

Wang, Z. (2016, 1). *Transactivation of epidermal growth factor receptor by g protein-coupled receptors: Recent progress, challenges and future research* (Vol. 17) (No. 1). Multidisciplinary Digital Publishing Institute (MDPI). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26771606http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4730337 doi: 10.3390/ijms17010095

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. New York: Wiley.

Yule, G. U. (1907). On the Theory of Correlation for any

Number of Variables, Treated by a New System of Notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *79*(529), 182–193. Retrieved from http://rspa.royalsocietypublishing.org/cgi/doi/10.1098/rspa.1907.0028 doi: 10.1098/rspa.1907.0028

Zhao, S., Shojaie, A., & Witten, D. (2017). In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *arXiv*(1), 1–61. Retrieved from http://arxiv.org/abs/1705.05543

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012, 4). The huge Package for High-dimensional Undirected Graph Estimation in R. *Journal of machine learning research : JMLR*, *13*, 1059–1062. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26834510http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4729207http://dl.acm.org/citation.cfm?id=2343681%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26834510%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcg doi: 10.1002/aur.1474.Replication