On Penalty Parameter Selection for Estimating Network Models

Anna C. Wysocki
University of California, Davis

Mijke Rhemtulla
University of California, Davis

Network models are gaining popularity as a way to estimate direct effects among psychological variables and investigate the structure of constructs. A key feature of network estimation is determining which edges are likely to be non-zero. In psychology, this is commonly achieved through the graphical lasso regularization method that estimates a precision matrix of Gaussian variables using an ℓ_1-penalty to push small values to zero. A tuning parameter, λ, controls the sparsity of the network. There are many methods to select λ, which can lead to vastly different graphs. The most common approach in psychological network applications is to minimize the extended Bayesian Information Criterion, but the consistency of this method for model selection has primarily been examined in high dimensional settings (i.e., $n < p$) that are uncommon in psychology. Further, there is some evidence that alternative selection methods may have superior performance. Here, with simulation, we compare four different methods for selecting λ, including the stability approach to regularization selection (StARS), K-fold cross-validation, the rotation information criterion (RIC), and the extended Bayesian information criterion (EBIC). Our results demonstrate that penalty parameter selection should be made based on data characteristics and the inferential goal (e.g., increase sensitivity versus avoidance of false positives). We end with recommendations for selecting the penalty parameter when using the graphical lasso.

Introduction

Network models are becoming more popular in psychology (Borsboom, 2017; McNally et al., 2015) largely because they provide a theoretical alternative to latent variable models and the common cause framework (Borsboom & Cramer, 2013; Schmittmann et al., 2013). For example, psychopathologies such as depression are often conceptualized as arising from a common or underlying cause. As this cause is unobservable (i.e., the latent variable), the symptoms are considered passive indicators that allow for inquiry and the ability to diagnose the disorder. On the other hand, networks conceptualize constructs as systems arising due to interactions between variables rather than due to an underlying cause (Epskamp, Maris, Waldorp, & Borsboom, 2017). In practice, network models are used to estimate relations between nodes (e.g., symptoms), identify hubs (i.e., highly connected nodes), and visualize the overall structure of a construct (McNally, 2016); for a theoretical discussion and comparison of latent versus network models see Borsboom and Cramer (2013).

One type of network is the Gaussian Graphical Model (GGM) wherein nodes represent random variables, and edges represent conditional independencies, estimated as partial correlations, between variables (Lauritzen, 1996). The estimation of partial correlations can produce rich inferences as variables that directly activate each other will be connected assuming all important variables are included in the model. As applied researchers can never be certain that this criterion is met, partial correlations provide a possible causal skeleton for a construct (Edwards, 2012). In psychology, GGMs are used to estimate, for example, symptom, personality, and health behavior networks (Costantini et al., 2015; Fried et al., 2017; Kossakowski et al., 2016).

An important aspect of GGMs is not only the identification of important conditional relations, but also the identification of truly zero edges, thereby achieving a sparse network. Setting edges to zero is a key feature of network estimation as, from a theoretical standpoint, having a fully connected model is less helpful than having a few potentially meaningful connections to focus on in future experiments or interventions. However, it is important to achieve this sparsity in a justified manner.

In psychology, inducing sparsity is typically done through a form of penalized maximum likelihood, the graphical lasso or "glasso" (Friedman, Hastie, & Tibshirani, 2008), using a penalty selection method called EBIC (Foygel & Drton, 2010) for selecting the degree to which the likelihood is penalized (for our purposes referred to as glasso$_{EBIC}$). Unlike traditional model selection with information criteria, which is explicitly used for edge selection, minimizing the EBIC is used to select the penalty parameter λ, that in turn achieves the goal of edge selection. Although there are multiple methods to select λ for the glasso equation (M. O. Kuismin & Sillanpää, 2017), glasso$_{EBIC}$ has emerged as the default in psychology[1]with no published work establishing if its performance is superior for psychological data. We understand that 'psychological data' is a broad term, and given the variety of research that is done in psychology (e.g., neural,

psychopathology, social research) no single dataset or template could characterize all of psychological data. However, networks estimated using the glasso have largely been applied to psychopathology and personality constructs (Beard et al., 2016; Briganti, Kempenaers, Braun, Fried, & Linkowski, 2018; Bryant et al., 2017; Pereira-Morales, Adan, & Forero, 2017). As such, when we use the term psychological data, we mean psychological data that networks have typically been fit to.

The selection of the penalty parameter term in glasso, λ, is critical as different values applied to the same data can result in different networks (Epskamp & Fried, 2018; M. O. Kuismin & Sillanpää, 2017). For example, when $\lambda = 0$ the network is no longer penalized, and, assuming no sample partial correlations are precisely zero, the resulting network is fully connected (i.e., all edges are non-zero). As λ increases, so does the penalization of the network, resulting in an increasingly sparse network (i.e., fewer edges are estimated) eventually resulting in an empty network (W. Liu, 2013). Within the glasso framework, penalty selection is done through an automated data mining process whereby a sequence of λs are tested, and one is selected based on whether the corresponding network optimizes some criterion. This criterion depends on which penalty selection method is being used. As different methods have divergent priorities (e.g., stability, sparsity, predictive ability), they often select different λs and can return vastly different networks (Kuismin & Sillanpaa, 2017). Characterizing the performance of penalty selection methods is the aim of the present work, in that we seek to fill this gap in the literature by comparing four penalty selection methods for the glasso in conditions representative of psychological data.

The rest of this paper is outlined as follows. In the next section, we describe network estimation. Then, we outline four penalty selection methods: CV, StARS, RIC, and EBIC, and discuss the advantage and limitations of each approach. Then, with a motivating example, we use each method to estimate the network structure of PTSD symptoms, where we highlight how each method can (sometimes) estimate drastically different networks. Importantly, we also provide an overview of specific network characteristics (e.g., network density and partial correlation size) based on a review of published psychological networks. We next present two simulation studies and their results ending with a discussion of these findings in relation to psychological networks, the practical implications of this work, and future directions for methodological inquiry.

Network Estimation

A key feature of estimating networks is imposing a sparsity pattern on the precision matrix ($\boldsymbol{\Theta}$),the inverse of the covariance matrix ($\boldsymbol{\Sigma}$). The sparsity pattern of Θ provides the structure of the network model where a non-zero value in the off-diagonal represents an estimated edge between nodes. The precision matrix can then be used to obtain partial correlations following

$$\text{cor}(y_i, y_j \mid y_{-(i,j)}) = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}},$$
$$\text{where } \boldsymbol{\Theta} = \begin{bmatrix} \theta_{ii} & & \\ \vdots & \ddots & \\ \theta_{ij} & \cdots & \theta_{jj} \end{bmatrix} = \boldsymbol{\Sigma}^{-1}. \tag{1}$$

As previously mentioned, in psychology, this is typically done through a regularization method called the graphical lasso, a form of penalized maximum likelihood using an ℓ_1-penalty (Friedman et al., 2008), which minimizes the likelihood:

$$l\left(\hat{\boldsymbol{\Theta}}\right) = \log\det\ \hat{\boldsymbol{\Theta}} - \text{tr}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}} - \lambda_p \sum_{i \neq j} \left(|\hat{\boldsymbol{\Theta}}_{i,j}|\right) \tag{2}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix and $\lambda_p \in [0, 1]$ is the penalty parameter controlling the penalization of the likelihood. The penalization occurs by setting a maximum limit on the sum of the absolute value of the edge weights. To meet this limit, a penalty term, λ, is applied to the sum of the absolute edge weights. Resulting in all edge weights being shrunk, many to zero. In other words, the glasso performs both regularization and edge selection impacting both which edges are estimated and the size of the estimated edge weights. When λ is set as 0, the penalty term drops off and the equation returns to regular (i.e., non-penalized) maximum likelihood:

$$l(\hat{\boldsymbol{\Theta}}) = \log\det\ \hat{\boldsymbol{\Theta}} - \text{tr}\left(\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}\right) \tag{3}$$

Again, there are different methods to select λ with divergent criteria. In the next section, we outline four penalty selection methods.

Methods for Parameter Selection

Cross-Validation (CV)

In psychology, there has been a surge of interest in predictive modeling, which stands in contrasts to more traditional explanation-centric frameworks (Yarkoni & Westfall, 2017). Here the goal is not exclusively to make inferences about individual parameters, but to select a model that is able to predict out-of-sample data. Of course, while inferences are still possible, some limitations exist such as cross-validation is prone to over-selection or a higher false positive rate (Chetverikov, Liao, & Chernozhukov, 2016; Yu & Feng, 2014). These limitations hold for cross-validation in both regression and network settings. Although there are different ways to implement cross-validation, the general

[1] glasso$_{EBIC}$ is the default in the most popular R package, *qgraph*, for fitting psychological network models

procedure is to partition the data into a training set and a test set. The training set selects a model, and the test set quantifies the prediction error of the training model. Different forms of CV (e.g., leave-one-out, K-fold) have been applied to network estimation (Efron, Hastie, Johnstone, & Tibshirani, 2004; Friedman et al., 2008; Friedman, Hastie, & Tibshirani, 2010; Zhang, 1993). For our purposes, we will focus on K-fold CV, which is computationally more efficient and has been found to have greater stability than other forms of CV (Homrighausen & McDonald, 2013, 2014).

K-fold CV partitions the data into K non-overlapping subgroups. Using Equation 2, a sparse precision matrix is estimated with the pooled data from $K-1$ of the subgroups, inverted into a sparse covariance matrix ($\hat{\mathbf{\Sigma}}_{\text{train}}$), and then tested on predicting the covariance matrix from the remaining group (i.e., the testing group; $\hat{\mathbf{\Sigma}}_{\text{test}}$). The prediction error is computed across folds (i.e., until each subgroup has been the test group) and averaged. The prediction error is estimated as

$$\ell_{CV}\left(\hat{\mathbf{\Sigma}}\right) = \frac{1}{K}\sum_{i=1}^{K} -\log\det\hat{\mathbf{\Sigma}}_{\text{train}} - \text{tr}\left(\hat{\mathbf{\Sigma}}_{\text{test}}\hat{\mathbf{\Sigma}}_{\text{train}}^{-1}\right). \quad (4)$$

This procedure is repeated across the range of λs resulting in a mean prediction error (prediction error is averaged across folds) for each λ. The λ that minimizes the mean cross-validation error is selected (for more details see Bien and Tibshirani 2011).

Rotation Information Criteria (RIC)

The RIC uses a rotation procedure to select λ (T. Zhao, Liu, Roeder, Lafferty, & Wasserman, 2012; Zhu & Cribben, 2018). Similar permutation or rotating techniques are regularly used in non-parametric tests wherein permuting the data by assigning different outcomes to each dependent variable observation constructs the expected sampling distribution given that the null hypothesis is true (i.e., the null distribution). The null distribution can then be used to evaluate the original data (Nichols & Holmes, 2003). The RIC randomly rotates the rows within each column creating a rotated dataset with only spurious relations between its variables. The RIC then finds the smallest value of λ that will accurately regularize all (spurious) edges to 0. This procedure is repeated a number of times (the default in its R package is 20) and the smallest calculated λ across rotations is returned as the selected penalty for the network.

Stability Approach to Regularization Selection (StARS)

StARS uses a re-sampling method to select a λ that provides maximal network stability (H. Liu, Roeder, & Wasserman, 2010). There are parallels between this method and bootstrapping (although bootstrapping re-samples with replacement and StARS without) wherein both methods use re-sampling to assess the variance or stability of a model across samples (Efron & Tibshirani, 1993). StARS does this by drawing K random, overlapping subsamples and fitting the range of λs to each of the subsamples. StARS begins with a large λ resulting in an empty network providing stability with no variation between groups and gradually reduces λ until there is a small but acceptable amount of variability between the subsample networks. Total instability for a given λ is defined as

$$D = \frac{\sum_{i\neq j}\left(2\left(\xi_{ij}(\lambda)\right)\left(1-\xi_{ij}(\lambda)\right)\right)}{\frac{p(p-1)}{2}} \quad (5)$$

where p is the number of variables in the dataset and $\xi_{ij}(\lambda)$ is the probability of a network having a specific edge calculated as

$$\xi_{ij} = \frac{\sum_{i\neq j}\hat{a}_{ij}}{K}, \quad (6)$$

where $\hat{a}_{ij} = 1$ when the corresponding element of $\hat{\mathbf{\Theta}}$ is nonzero and K is the number of subsamples. The smallest λ with a total instability between $.01 < D(\lambda) < .08$ is selected (H. Liu et al., 2010).

Extended Bayesian Information Criterion (EBIC)

Information criteria, for the purpose of model section, are commonly used in psychology, and most can be justified in several ways, for example, minimizing the BIC approximates selecting the most probable model, assuming the true model is in the candidate set (Raftery, 1995). When the ratio of sample size to number of variables is small, it was noted that the Bayesian information criterion does not necessarily select a parsimonious model (Chen & Chen, 2008). As such, EBIC was developed by introducing an additional manually set penalty, γ, to the BIC equation (Foygel & Drton, 2010). γ controls the prior probability of sparse models resulting in the return of sparser networks as γ increases (Chen & Chen, 2008, 2012). EBIC is calculated as

$$\text{EBIC} = -2l(\hat{\mathbf{\Theta}}) + E\log(n) + 4\gamma E\log(p), \quad (7)$$

where $l(\hat{\mathbf{\Theta}})$ is defined in Equation 2 and E is the size of the edge set (i.e., the number of non-zero elements of $\hat{\Theta}$). When $\gamma = 0$ the added penalty is dropped, and the equation is reduced to the BIC. The selected network minimizes the EBIC with respect to λ. Per recommendations by Foygel and Drton (2010), the default setting for γ in popular R packages for estimating network models, such as qgraph, glasso, and huge, is .5 (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; Friedman, Hastie, & Tibshirani, 2018; T. Zhao et al., 2012).

Epskamp, Borsboom, and Fried (2018) investigated EBIC's performance across conditions ($p = 25$, $50 \leq n \leq 2500$) that

are typical to psychology and showed that the sensitivity (defined in section Performance Measures) of EBIC increased with sample size. Specifically, when sample size was less than 500 sensitivity was generally below 75% (i.e., the method only estimates 75% of the true edges). Additionally, EBIC seemed to be sensitive to its population network's partial correlation size wherein larger partial correlations tended to negatively impact method performance (Williams, Rhemtulla, Wysocki, & Rast, 2018).

Method Performance

The glasso has a number of assumptions and necessary conditions on which its performance depends. One assumption is that the underlying population matrix is sparse. In the statistics literature, sparsity can be defined as having fewer true edges than sample size (Meinshausen & Bühlmann, 2006; Tibshirani, 2015). This is the minimum level of sparsity required for the network to be estimable. However, in psychology, sample sizes generally greatly exceed the number of variables, so the number of possible edges rarely meets this limit (see Section Psychological Network Review). As such, it is unclear what the impact of sparsity will be in this case. Second, there are two conditions that must be met to ensure consistent estimation. The irrepresentable condition is satisfied when unimportant variables are not highly correlated with important variables. More specifically, this condition is satisfied when the sum of irrelevant covariance is less than 1 (P. Zhao & Yu, 2006). The beta-min condition is satisfied when non-zero coefficients are sufficiently large; see Gauraha and Swapan (2018) for a full discussion of these necessary conditions.

Further, as these methods were developed for data where the ratio of sample size (n) to the number of variables (p) is small, most simulations have explored glasso's performance in such high-dimensional settings (Foygel & Drton, 2010; Friedman et al., 2010). Of the simulations that have looked at glasso's performance in low-dimensional settings, where p is smaller than n, many have only used EBIC to select the penalty parameter (Epskamp, 2016; Epskamp & Fried, 2018; Williams et al., 2018). However, there is a small amount of work comparing penalty parameter selection methods. For example, StARS was found to be competitive against BIC and AIC in conditions with a small n to p ratio (performance was measured using F_1-scores, a measure of both recall and precision; H. Liu et al. 2010). Overall, StARS consistently estimated sparser graphs than other methods and as a result had a lower false positive rate but was also less sensitive (see Section Performance Measures for definition). K-fold CV was also compared to StARS and was found to return a denser network with more false positive errors but fewer false negative errors (H. Liu et al., 2010). Mohammadi and Wit (2015) compared EBIC, RIC, and StARS across four low-dimensional conditions. The results suggest RIC and StARS were competitive with EBIC even outperforming (performance again measured by F_1-scores) EBIC in multiple conditions particularly those most comparable to psychological data.

In simulations comparing penalty selection methods, there have been few low-dimensional conditions. This means that there has not been a full characterization of the impact of sample size and variable number across methods. It also may be important to characterize these methods specifically for psychology-typical data. GGMs have typically been used to estimate gene or neural networks (Hecker, Lambeck, Toepfer, van Someren, & Guthke, 2009). However, in psychology they are being used to estimate constructs such as symptom or personality networks. As these are vastly different areas, there is the possibility that psychological networks contain different sized partial correlations than gene and neural networks. Given the impact partial correlation size seems to have on these methods' performance, it is important to characterize method performance with partial correlations that approximate the size found in psychology. Finally, to our knowledge, no simulation comparing penalty selection methods has directly manipulated sparsity, the percentage of truly zero edges, across conditions. Therefore, we do not know whether sparsity or the lack thereof impacts penalty selection methods differently. As such, it is important to assess how these methods perform in conditions and data that are typical to psychology.

Psychological Network Review

To better understand the characteristics of psychological networks, we reviewed 37 recently published psychological networks assessing psychopathology (n = 33) and personality (n = 4) constructs. As the data were similar between these two constructs, their results will be presented together. Our review focused on the sample size, number of variables, sample sparsity and estimated partial correlations for each network. Figure 1A depicts the distribution of partial correlations. Note these are non-regularized partial correlations (i.e., unbiased estimates). From this figure, we can see most networks have many partial correlations near 0, and large partial correlations are less frequent across networks. It is likely that these small partial correlations would be set to 0 through regularization. Although small partial correlations are more frequent, most networks have moderate to large partial correlations as well (see Table 1 for median partial correlation ranges and Table 2 for percentage of partial correlations within ranges).

To obtain an estimate of sparsity without the use of a regularization method, we set any partial correlations less than .05 to 0. Figure 1B depicts the sparsity distribution across networks. Network sparsity varied from 25 to 75%. 76% of the networks had between 50 and 75% sparsity. Table 2

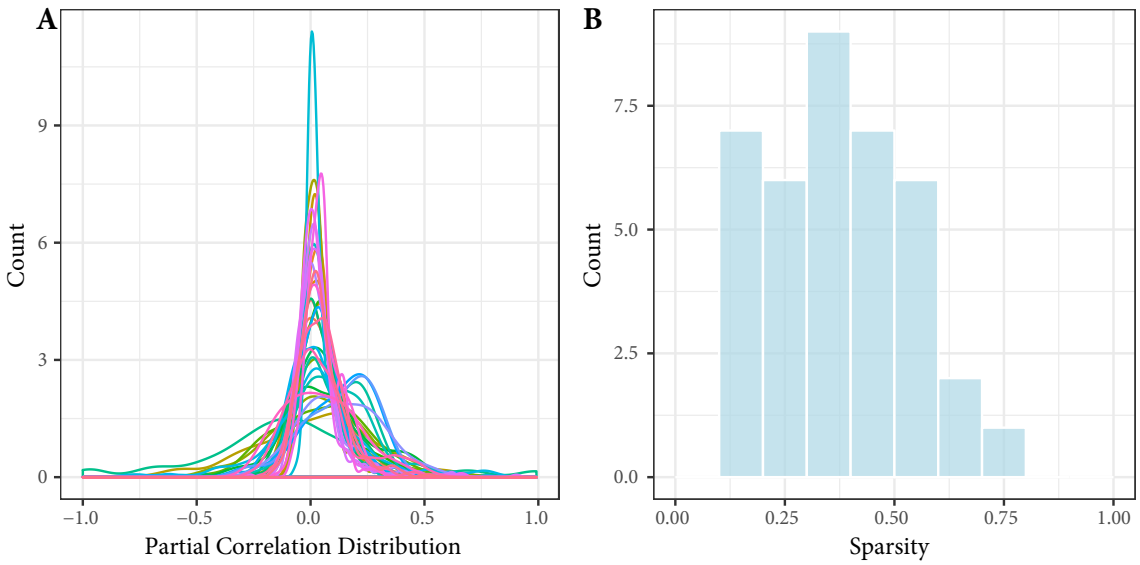

Figure 1. A: The distribution of estimated partial correlations across 37 networks. Each line represents a different network. Partial correlation size is depicted on the x-axis. B: The distribution of sample sparsity across 37 networks. Sparsity is depicted on the x-axis.

Table 1

	Median	Mean	SD
Sample Size	404.5	1,044.74	2420.04
Variables	18.5	19.84	7.25
Max PC	.52	.56	.18
Min PC	.000	.001	.003

displays the percentage of networks that fall between specific ranges for the features sparsity, sample size, and variable number (see Figure A1 in the supplementary material for the distribution of sample size and variables across the networks). In this table, we can see 55% of the networks had a sample size less than 500, and most networks had between 10 and 30 variables. From this review, we surmise that psychological networks are often not extremely sparse and contain many small and a few larger partial correlations. Further, sample sizes are typically under 500, and the number of variables ranges between 10 to 30.

Motivating Example

To highlight the differences between methods, we estimated four network models, each using a different penalty selection method, from a post-traumatic stress disorder (PTSD) data set (McNally, 2016) Each method selected a different λ and estimated networks with different sparsities, although EBIC and RIC had near identical sparsities (< 1% difference; See Figure 1A for visualization of each network). Even if a similar sparsity level is estimated across methods, λ also impacts the edge weights (see Figure 1B). Comparing EBIC and RIC, even though each network has a similar number of edges, the edge weights within the RIC selected network are smaller. Further, CV not only selects the densest network but also estimates the largest absolute edge weights while StARS selects the sparsest network with the smallest absolute edge weights. This example underscores λ's impact on which edges get estimated and the edge weights, and, given that each penalty selection method selects a different lambda it bolsters the need for guidance on which method to use.

Simulation Study

In psychology, no recommendations exist for the use of one penalty selection method over another. Rather the performance of one method, glasso$_{EBIC}$, has been examined across various conditions (Epskamp, 2016; Epskamp & Fried, 2018). In an effort to bridge this gap, we conducted two simulation studies. The first used partial correlations estimated from a psychological dataset (Section Simulation 1: Empirical Partial Correlations), whereas the second simulated data to assess the effect of partial correlation size on performance (Section Simulation 2: Simulated Partial Correlations). By using empirical partial correlations from a psychological dataset we are establishing how these methods may reasonably perform when used to fit a psychological network. Then, by varying partial correlation size, still within a range that is representative of psychological data, we are establishing how performance may vary with respect to different partial correlation ranges. The code for the both simulations can be

[2]Based on the absolute values of the partial correlations

Table 2

Sparsity		Partial Correlation[2]		Sample Size		Number of Variables	
Range	Percentage	Range	Percentage	Range	Percentage	Range	Percentage
0 - .25	0%	0 - .15	74.1%	100 - 250	42%	0 - 10	18%
.25 - .50	24%	.15 - .25	13.5%	250 - 500	13%	10 - 20	47%
.50 - .75	76%	.25 - .50	9.7%	500 - 1,000	26%	20 - 30	24%
.75 - 1	0%	.50 - 1	2.7%	>1,000	18%	>30	11%

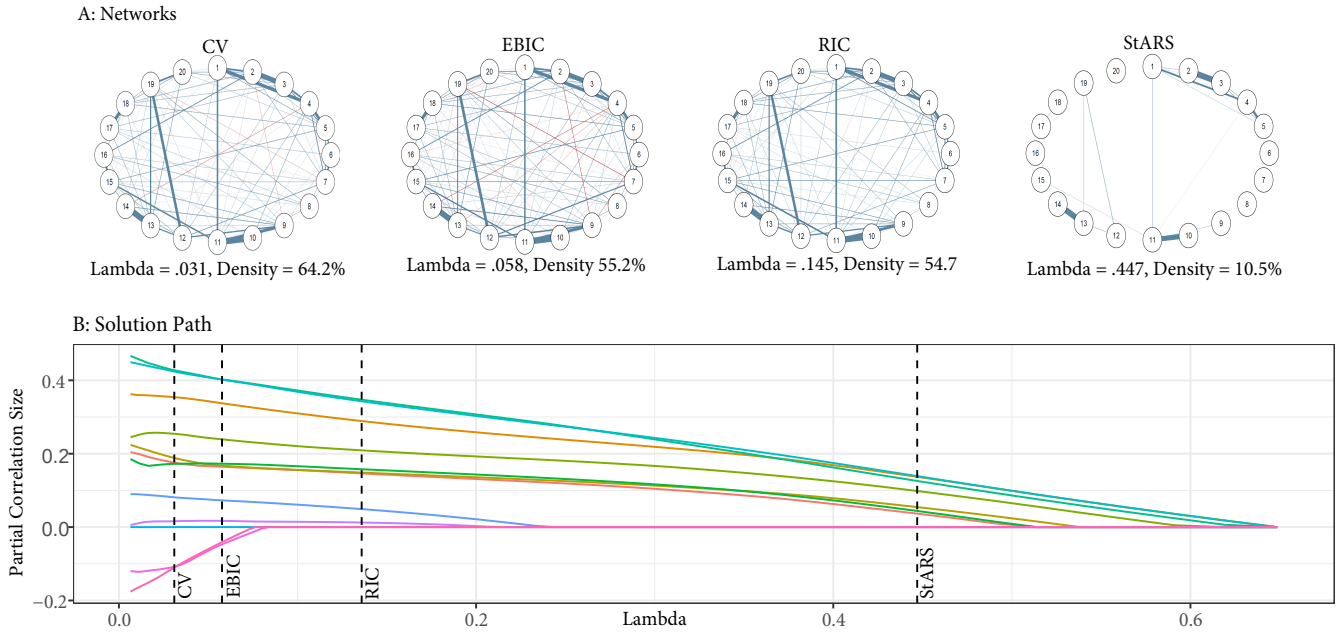

Figure 2. A: Returned networks estimated from a psychological dataset for each method. B: the solution path for each method. Each line depicts an edge (12 were selected out of the edge set) and the change in partial correlation size across lambdas. The dashed line represents the selected lambda for different methods.

found at https://github.com/AnnaWysocki/Network-Penalty-Selection

Simulation 1: Empirical Partial Correlations

In Simulation 1, we used a 20 variable PTSD symptom dataset(McNally et al., 2015). We selected this dataset for two reasons. First, the data have previously been used to assess the performance of glasso$_{EBIC}$ allowing for greater comparability between our results and previous simulations, and, second, simulating data based on a psychopathology dataset provides partial correlations that are comparable in size to those that are likely to appear in psychological network applications (see Figure A2 in the supplementary material for the distribution of estimated partial correlations).

Our goal was to outline how these methods compared to each other along with how data characteristics such as sparsity, number of variables, and sample size influenced their performance. In addition, preliminary simulations suggested that the performance of these methods varied not only across samples and conditions but also across populations within a condition (e.g., two networks with the same level of sparsity, number of variables, and sample size with a different population matrix may return different false positive rates, on average, across randomly sampled data). To better assess this within-condition variability, we performed a nested model simulation to estimate both sampling variability within populations and between-population variability within conditions (see Figures A5 and A6 in the supplementary materials). In other words, we simulated multiple datasets from the same population matrix as well as multiple populations within each simulation condition (described in greater detail below). We varied the number of variables (p; 10 and 20), sparsity level (50% and 80%), and sample size (n; 100, 200, 250, 500, 1,000, 2,000, and 3,000) across conditions.

The simulation procedure to create each population matrix was as follows:

1. We used the PTSD dataset to form a bank of partial correlations
2. All partial correlations within the range of ± 0.05 were removed (following Epskamp 2016, simulation procedure)
3. X partial correlations were randomly sampled, without replacement, from the bank to create the population partial correlation matrix where

$$X = \left(\frac{p\,(p-1)}{2}\right) * (1 - \text{sparsity}), \tag{8}$$

$p \in \{10\ ,\ 20\}$, and sparsity $\in \{.50\ ,\ .80\}$

Thus, we had 2 (sparsity levels) by 2 (p) by 7 (n) conditions (i.e., 28 conditions). For each condition, 100 population matrices were created. For each population matrix we carried out the following procedure:

1. Simulate 1,000 multivariate normally distributed datasets of size n
2. With each dataset, estimate four networks using the four previously outlined penalty selection methods
3. Compute performance measures (see Section Performance Measures).

Simulation 2: Simulated Partial Correlations

Simulation two assessed the effect of partial correlation size on method performance and more fully characterized the effect of sparsity. The edge weights were randomly generated from a G-Wishart distribution which is frequently used to simulate multivariate data as it can be defined with only two parameters $\mathbf{\Theta} \sim W_G(df, I_p)$ where I_p represents a p by p identity matrix (Mohammadi & Wit, 2017). The degree of freedom parameter determines the degree of shrinkage towards the identity matrix. As the parameter increases the distribution of θ_{ij} narrows. In other words, as the degree of freedom parameter increases the partial correlations approach zero (Hsu, Sinay, & Hsu, 2012) through reduction of tail-heaviness (i.e., fewer extreme values). We adjusted the degrees of freedom to correspond to two ranges where 90% of the partial correlations on average fell between ±.35 and between ±.25.

Simulating from a G-Wishart distribution also guarantees a positive definite posterior estimate for $\mathbf{\Theta}$ (M. Kuismin & Sillanpää, 2016). To achieve a positive definite matrix, the partial correlations must become smaller as the network becomes more connected. As such, the same degree of freedom will result in smaller partial correlations when sparsity is at 50% compared to when it is at 80%. To account for this, we determined which degrees of freedom corresponded to the previously mentioned partial correlation ranges for each level of sparsity. Sparsity varied from .1 to .9 in increments of .2. Sample size conditions were 100, 200, 250, 500, 1,000, 2,000, and 3,000, as in Simulation 1, and p was fixed at 20. In total, we had 70 conditions.

Performance Measures

We were interested in quantifying the performance of these methods with respect to both the accuracy of edge detection and the accuracy of edge weights. For edge detection we calculated the sensitivity (i.e., the true positive rate) and the false positive rate (FPR) as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ and } \text{FPR} = 1 - \left(\frac{\text{TN}}{\text{TN} + \text{FP}}\right). \tag{9}$$

where TP is the number of true positives (i.e., the number of estimated edges that are non-zero in the population matrix), FN the number of false negatives (i.e., the number of unestimated edges that are non-zero in the population matrix), TN the number of true negatives, and FP the number of false positives detected by each method. Sensitivity and the FPR range from 0 to 1. A sensitivity score of 1 indicates that the method is correctly detecting all true positives, and a FPR of 0 means the method is correctly estimating all true negatives as exactly zero. A method with perfect edge detection would have have a sensitivity score of 1 and FPR of 0. Finally, we compared the sparsity of the estimated network to population sparsity to assess whether the methods were sensitive to population sparsity.

A method with perfect edge detection (i.e., the method is estimating all true positives and no false positives) may still estimate edge weights incorrectly. To capture the accuracy of the estimated edge, we calculated the correlation between the non-zero partial correlations in the population and the corresponding estimated edge weights. We refer to this performance measure as true edge correlation. Note, if an edge was non-zero in the population matrix but zero in the estimated matrix (i.e., a false negative) it was included in the estimation of the true edge correlation as we were interested in the correlation between the estimated and population values of *true* edges.

Results

Simulation 1

Sensitivity and FPR are presented in Figure 3. The performance measure is denoted on the y-axis. The columns and rows correspond to different simulation conditions (p and sparsity), and sample size is denoted on the x-axis. The four methods are depicted as different lines, and the shading around the lines represents ± one standard deviation of the outcome. The variability depicted in the figures represents

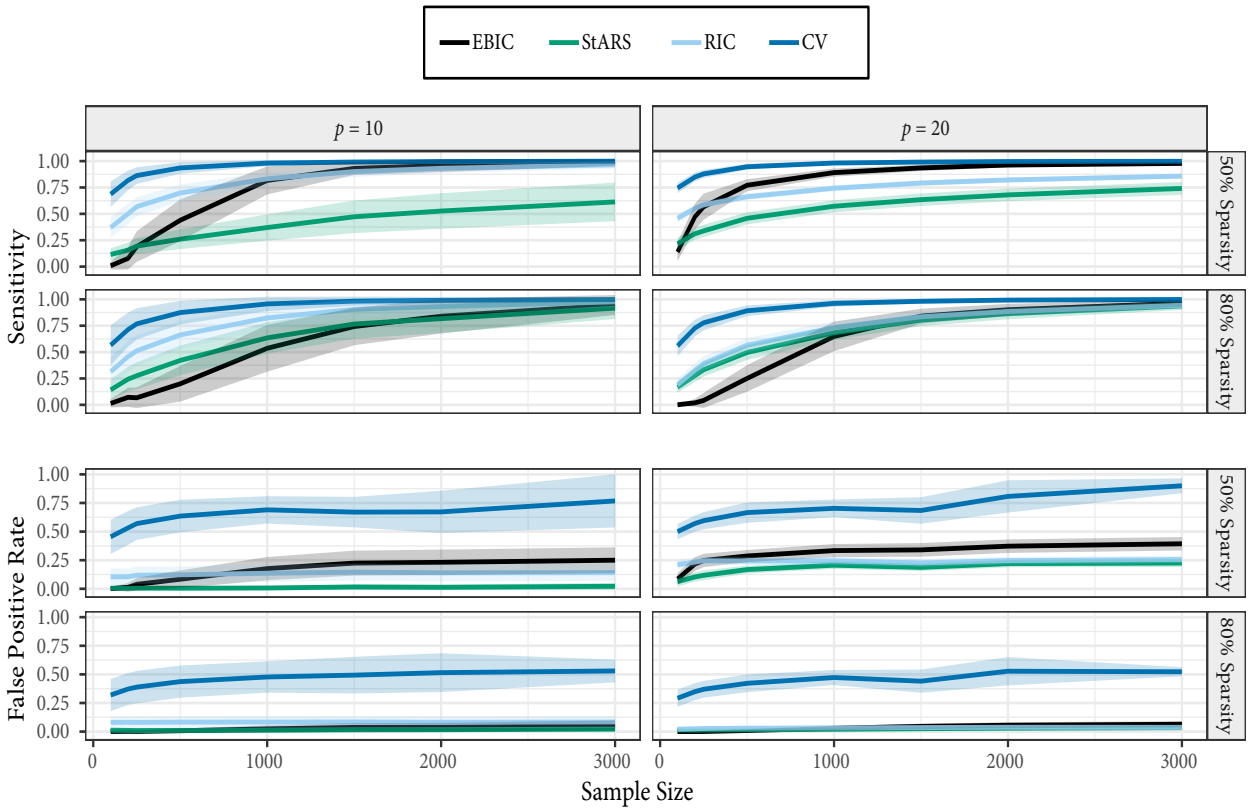

Figure 3. Sensitivity and False Positive Rate for four methods (CV, EBIC, RIC, and StARS) simulated using a psychological dataset. The columns (from left to right) correspond to the number of variables p (10 or 20), and the rows correspond to population sparsity (i.e., the percentage of edges equal to 0 in the population; 50% or 80%). The shading around each of the lines represents the average sampling variability within a population.

the average sampling variability within a population. Figure A5 in the supplementary material depicts the between-population variability.

For smaller sample sizes (n < 500), EBIC often returns empty networks. This can be improved by decreasing the γ parameter. However, we set γ = .5 for all conditions per recommendations in the psychological network literature (Epskamp & Fried, 2018). As such, a number of networks within the $n \in \{100\ ,\ 200\}$ conditions were empty (particularly n = 100). The other methods never returned an empty network. However, StARS and RIC in those same conditions often returned exceptionally sparse networks (i.e., < 10% of possible edges were estimated).

Sensitivity. Across both levels of p and both levels of sparsity, CV had far higher sensitivity than the other three methods. For sample size 500 and above, CV consistently had sensitivity rates of greater than 80%. EBIC, RIC, and StARS, in contrast, had low sensitivity at small n. For these three methods, although sensitivity improved as sample size increased, it is did not reach 80% until sample size was greater than 1,000. For many conditions with a sample size of 500 or less, sensitivity is under 50% particularly for StARS and EBIC. Sparsity did not have a great impact on sensitivity, but the size of the network did. When p = 10, all methods had higher sampling variability compared to p = 20, meaning the performance of a method was more uncertain. This was particularly true for EBIC. Greater variability when p = 10 was not unexpected as the performance metrics are computed as a proportion of the total number of edges. As such, when there are fewer potential edges each single edge has a greater influence on both the mean and variability of a performance index.

FPR. The FPR results largely mirrored the sensitivity results: methods and conditions with higher (better) sensitivity typically exhibited higher (worse) FPRs. CV had a markedly higher FPR across all conditions compared to the other three methods. Unlike sensitivity, the FPR of all methods was impacted by sparsity. Specifically, there was an interaction between sparsity and sample size. Within the sparse condition (sparsity = 80%), the FPR was generally low and constant across sample size. Once sparsity decreased to 50% the FPR was higher and increased as sample size increased. For example in the p = 20 and n = 500 conditions, EBIC on average has an FPR of 5% when sparsity is 80% but

the FPR rate increases to 60% when sparsity is 50%. The effect of sample size on the FPR may be explained by the decreased penalization of $\hat{\Theta}$ (i.e., smaller λs are being selected; see Figure A3 in the supplementary material for a visualization of selected lambda across sample size) as sample size increases, resulting in a denser estimated matrix. Note that StARS and RIC are more robust to this interaction. Like sensitivity, network size affected the variability of the results, but not the trends.

True Edge Correlation. The results for true edge correlation were similar to those for sensitivity (i.e., low correlation when $n > 1000$, greater variability at $p = 10$). Due to this similarity, all True Edge Correlation figures are included in the supplementary material (see Figure A4 in the supplementary material).

Simulation 2

Sensitivity and FPR are presented in Figure 4. Note, since we had five levels, sparsity is now denoted on the x-axis. The rows correspond to sample sizes. We selected three characteristic levels of sample size (250, 500, and 1000) to visualize, and the columns correspond to the average partial correlation ranges (±.35, ±.25) that contain 90% of the values. Each line depicts a different method, and the shading around the lines represents ± one standard deviation of the outcome.

Sensitivity. In agreement with Simulation 1, CV has the highest sensitivity across all conditions compared to the other three methods, and the sensitivity of all methods increases with sample size. Sparsity had an impact on methods in specific conditions but this effect was not consistent in direction across methods or conditions. For example, with larger sized partial correlations and a sample size of 1,000 sensitivity increases with sparsity for StARS and RIC but decreases for EBIC. In this condition, CV is not impacted. However, we noted a large effect of partial correlation size on sensitivity wherein conditions with larger partial correlations had higher sensitivity, particularly when sparsity is low. This is likely because larger effects require less power for detection.

FPR. A number of the results from Simulation 2 were in agreement with Simulation 1. First, there was a trade-off between sensitivity and FPR wherein conditions with high sensitivity also had a high FPR, and those with less sensitivity tended to also have a lower FPR. This extended to methods more generally wherein methods with high sensitivity tend to also have a high-false positive rate (e.g., CV). Second, CV had a higher FPR across all conditions in comparison with the other methods. Third, both sparsity and sample size had an impact on FPR. Specifically, the FPR increased as sparsity decreased and sample size increased.

Simulation 2 allowed us to investigate the effect of partial correlation size. Importantly, the effect of sample size and sparsity on FPR was moderated by the absolute partial correlation size. Notably, networks with smaller absolute partial correlations had a lower FPR across all conditions compared to those with larger absolute partial correlations. When on average 90% of the partial correlations were between ± 0.25, sample size and sparsity had less impact on the FPR, particularly for RIC and StARS. However, as the average partial correlation range increases, so does the FPR, and the interaction between sample size and sparsity is greater. Although this is the case for all methods, it is most apparent for EBIC and CV.

To explain the effect of partial correlation size, it is important to consider the assumptions and necessary conditions inherent in the graphical lasso (for more details see the Discussion). For example, consistent model selection depends on the irrepresentable condition being fulfilled (Meinshausen & Bühlmann, 2006). This condition is more likely to fail as sparsity decreases. When this condition is not met, the true edges are excessively penalized, and more false positives are estimated increasing the false positive rate (P. Zhao & Yu, 2006).

True Edge Correlation. True edge correlation (depicted in Figure A7 in the supplementary material) does not seem to be greatly impacted by sparsity. However, conditions with larger partial correlations tend to have higher true edge correlations. This is most pronounced for EBIC, but observed across all methods.

Population versus Estimated Sparsity. Figure 4 depicts population sparsity on the x-axis and estimated sparsity on the y-axis. The rows represent sample size, and the columns represent different methods. The dashed line represents the accurate estimation of sparsity for reference (i.e., when the population and estimated network's sparsity are equal). Note that, RIC and StARS tend to underestimate the number of edges, although it does improve as sparsity and sample size increases. CV tends to overestimate the number of edges. EBIC tends to overestimate edge number in conditions with larger partial correlations, and underestimate in conditions with smaller partial correlations. Importantly, EBIC has high accuracy when sample size is large and and the partial correlations are smaller. However, it is important to take into consideration the FPR. Even though there are conditions where the methods estimate the correct *number* of edges, it is likely they are estimating many of the wrong edges.

Overall, it appears like these methods are often insensitive to population sparsity. Given the impact of sparsity on method performance, it is important for researchers using the glasso to consider the population sparsity of the construct they are investigating (e.g., do I expect this construct to have many true connections?). However, the results comparing population to estimated sparsity show that estimated sparsity cannot be used to infer population sparsity. For example, a researcher cannot assume since the estimated

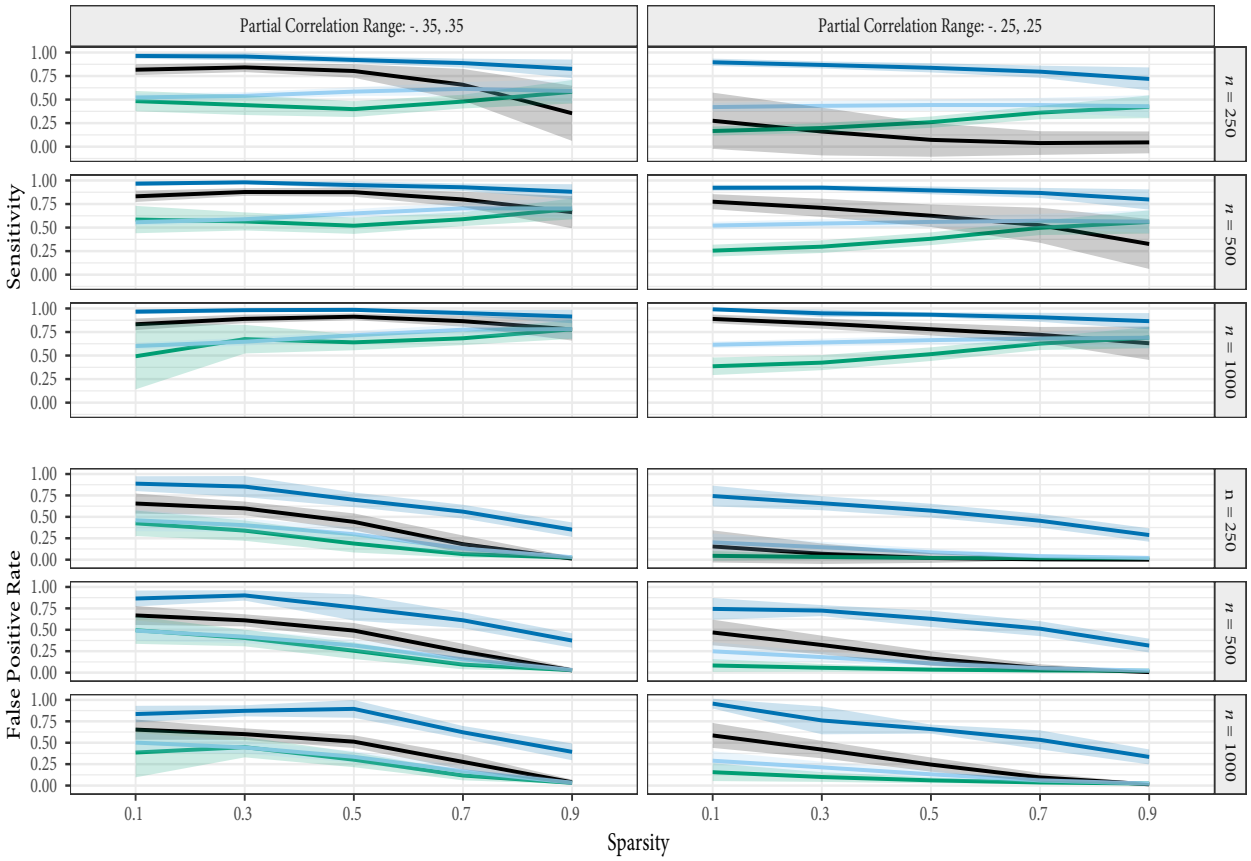

Figure 4. FPR and Sensitivity across sparsity levels. Lines denote different methods. Columns represent different partial correlation ranges, and rows represent sample size.

network has 30% sparsity the population network must also be sparse. This is particularly true as these methods do not consistently under or over-estimate sparsity. Rather the size of the population partial correlations, which is also unknown to applied researchers, impacts whether sparsity is under, over, or accurately-estimated.

Discussion

In this paper, we compared four methods for selecting the penalty parameter to compute the graphical lasso across conditions that are typical in psychology. We found that all methods had concerning performance features. First, all methods, with the exception of cross-validation, had low sensitivity at sample sizes that were most typical to psychology (n < 1,000). Second, we found that the sparsity of the population network, absolute partial correlation size, and sample size all impact the false positive rate. Notably, for networks with larger absolute partial correlations, there was an interaction between sparsity and sample size wherein the FPR increased as sample size increased, particularly when the population network was dense. This interaction was not observed or was attenuated for networks with smaller absolute partial correlations. Third, we found greater variability both within and between conditions and populations as sparsity decreased. This is important as it suggests that as sparsity decreases the uncertainty of the results increases. Finally, we found a consistent trade-off between sensitivity and the FPR when comparing both methods and conditions. For example, cross-validation had extremely high sensitivity (a desirable feature) and a high false positive rate (an undesirable feature). Additionally, conditions with larger absolute partial correlation sizes had high sensitivity but also a high false positive rate. The opposite was found for conditions with smaller absolute partial correlations.

When considering these results, it is important to note that regularization methods, including the glasso, were developed for high-dimensional situations (i.e., $n << p$). In these cases, the inverse of the sample covariance matrix cannot be computed due to singularity: $\det(\mathbf{\Sigma}) = 0$. Regularization, by shrinking the variable space, allows for the estimation of a high-dimensional matrix. However, to accurately estimate the underlying network the assumption of sparsity and the irrepresentable and beta-min (i.e., non-zero coefficients must be sufficiently large) conditions must hold (Meinshausen &

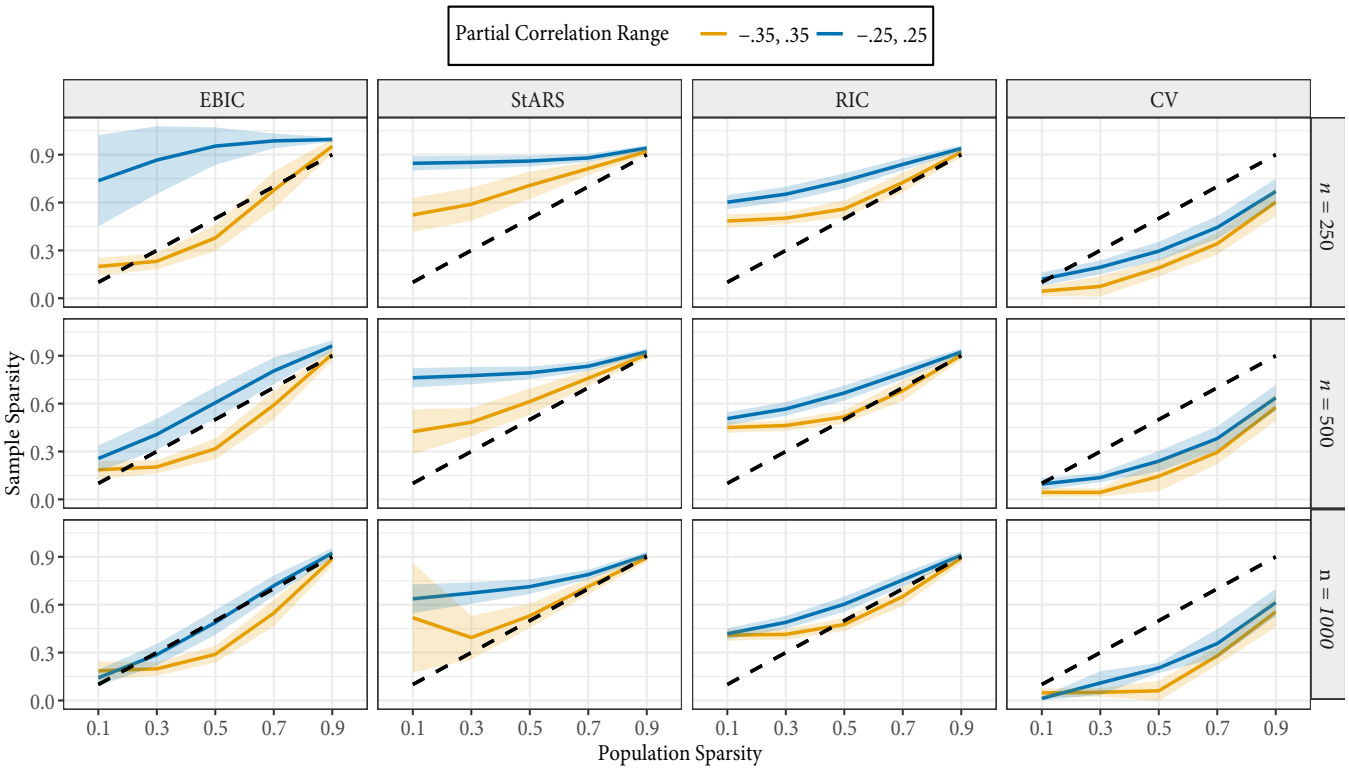

Figure 5. Population versus estimated sparsity. The dashed line denotes the correct level of sparsity (i.e., equal sparsity between population and estimated sparsity). Lines represent different partial correlation ranges wherein on average 90% of the partial correlations are between the values (e.g., -25, .25)

Bühlmann, 2006; P. Zhao & Yu, 2006). Previous research has characterized method performance when these conditions are met (Fu & Knight, 2000), but these conditions have been found to rarely hold without explicit specification (P. Zhao & Yu, 2006). In our simulations, we did not ensure these assumptions were met. Rather we were guided by the sparsity and partial correlation sizes observed in psychological applications of network models.

Our results demonstrate that as sparsity decreases (i.e., the population network increases in density) the FPR is higher and there is more variability in method performance both within and across conditions. This suggests that the methods are not robust to the violation of sparsity, and the population sparsity of a matrix is important to consider before using these methods. However, it is important to note that simply because a sparse matrix is estimated does not necessarily mean the population matrix is sparse as none of the methods included in our simulations were consistently sensitive to population sparsity. Our results also indicate when there are larger absolute partial correlations in the population network the FPR increases. One potential explanation is the failure of the irrepresentable condition which can result in an increase in false positives due to an over-penalization of true edges (P. Zhao & Yu, 2006). Overall, these results strongly suggest that researchers think about the possible properties of the construct they are investigating to qualify whether a regularization method would be appropriate for their data (e.g., is it likely the population network is sparse?). This is important as there are many situations, particularly in low-dimensions, where the costs of regularization supersede the benefits (Williams & Rast, 2018; Williams et al., 2018). It is also important to note that some of these performance issues can be attenuated. For example, sensitivity can be improved with a large sample size, n > 1,000, and the FPR can be improved with the use of thresholding (i.e., setting smaller edges to 0 after regularization) which has recently been suggested in psychology (Epskamp, 2018). When deciding on which penalty parameter selection method to use, it is important to consider the research goal. Given the consistent trade-off observed between sensitivity and the FPR, an important consideration is whether controlling the FPR or increasing sensitivity is more important to the research question at hand. If the goal is to obtain a dense network that contains the true model, then cross-validation, due to its high sensitivity, may be an appropriate method. In this case, a high FPR is acceptable. Across our simulations, StARS consistently returned the sparsest network so if the research goal is to return a network containing only the most important

Table 3
Method Strengths and Weaknesses

	Strength	Weakness
EBIC	Moderately High Sensitivity	Inconsistent Results
CV	High Sensitivity	High False Positive Rate
StARS	Low False Positive Rate	Low Sensitivity
RIC	Low False Positive Rate	Moderate Sensitivity

edges, then StARS, due to its low FPR, may be an appropriate method. In this case, low sensitivity is acceptable (see Table 3 for summary of each method's strengths and weaknesses). Finally, if both sensitivity and the FPR are equally important then using the RIC may be best as it tends to return the most balanced network (RIC had the highest F1-score, a measure of both recall and precision). In other words, a RIC selected network would tend to have more sensitivity than StARS but less than CV and a lower FPR than CV but higher than StARS.

Another consideration for penalty method choice is whether the researcher is most interested in prediction or explanation. For example, a cross-validation method may be appropriate when the goal is prediction. However, when the goal is explanation, CV's high FPR may be prohibitive even though it is highly sensitive.

In summary, EBIC seems to be greatly influenced by the characteristics of the population and data, more so than RIC and StARS, such as sparsity, sample size, and partial correlation size. Although RIC and StARS are influenced by these factors as well, it is to a lesser extent compared to EBIC. Further, RIC tends to strike the best balance between sensitivity and the FPR, compared to the other methods. As such, if the glasso is being used to fit a psychological network, using the RIC to select λ may be more appropriate.

Limitations and Future Directions

There are several important limitations of the present research. First, although our purpose was to assess which penalty selection method would be best suited for psychological data, there are many more penalty selection methods than the ones compared here (see M. O. Kuismin and Sillanpää 2017 for an over-view of different penalty selection methods). As such, our recommendations cannot be generalized to all penalty parameter selection methods. We chose the methods included in this paper based on previous research demonstrating they were competitive to the default method in psychology (H. Liu et al., 2010; Mohammadi & Wit, 2015), glasso$_{EBIC}$, and based on their ease of implementation for applied researchers (i.e., the availability of an R-package or code to fit these models). Future research may want to extend these simulations to include more penalty selection methods. On a similar note, we did not consider other methods of regularization such as Lasso regression which has been found to have promising performance even in low-dimensional settings (Mohammadi & Wit, 2015). Further research is needed to assess whether other forms of regularization could outperform the graphical lasso in these settings.

Conclusion

In conclusion, network modeling is an important tool in psychological research. However, the estimation techniques used are new to psychological data and as such should be used thoughtfully and cautiously. Although, all methods included in our simulations had some concerning performance, based on our results, we found that RIC had the most balanced performance. Importantly, our results underscore the need to carefully consider the likely population properties and the inferential goal before implementing a penalty selection method.

References

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M., … Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, *42*(2), 407–420. doi: 10.1002/jmri.24785.Free-Breathing

Bien, J., & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, *98*(4), 807–820. doi: 10.1093/biomet/asr054

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. doi: 10.1002/wps.20375

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

Briganti, G., Kempenaers, C., Braun, S., Fried, E. I., & Linkowski, P. (2018). Network analysis of empathy items from the interpersonal reactivity index in 1973 young adults. *Psychiatry Research*, *265*(September 2017), 87–92. Retrieved from https://doi.org/10.1016/j.psychres.2018.03.082 doi: 10.1016/j.psychres.2018.03.082

Bryant, R. A., Creamer, M., O'Donnell, M., Forbes, D., McFarlane, A. C., Silove, D., & Hadzi-Pavlovic, D. (2017). Acute and chronic posttraumatic stress symptoms in the emergence of posttraumatic stress disorder a network analysis. *JAMA Psychiatry*, *74*(2), 135–142. doi: 10.1001/jamapsychiatry.2016.3470

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces model selection. *Biometrika*, *95*(3), 759–771. Retrieved from http://www.jstor.org/stable/20441500%0Ahttp://www.jstor.org/stable/ doi: 10.1093/biomet/asnO34

Chen, J., & Chen, Z. (2012). EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM. *Statistica Sinica*, *22*, 555–574.

Chetverikov, D., Liao, Z., & Chernozhukov, V. (2016). On Cross-Validated Lasso. *arXIV preprint*.

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

Edwards, D. (2012). *Introduction to graphical modeling*. Springer Science & Business Media.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, *32*(2), 407–499. Retrieved from http://statweb.stanford.edu/~tibs/ftp/lars.pdf

Efron, B., & Tibshirani, R. (1993). An Introduction To The Bootstrap. *Journal of the American Statistical Association*, *89*, 436. doi: 10.1007/978-1-4899-4541-9

Epskamp, S. (2016). Brief Report on Estimating Regularized Gaussian Networks from Continuous and Ordinal Data. *arXIV preprint*. Retrieved from http://arxiv.org/abs/1606.05771

Epskamp, S. (2018, 5). *New features in qgraph 1.5 [Blog post]*. Retrieved from http://psychosystems.org/qgraph_1.5

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. (2018). *A Tutorial on Regularized Partial Correlation Networks*. doi: 10.1037/met0000167

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2017). Network psychometrics. In *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. 2-2, pp. 953–986). doi: 10.1002/9781118489772.ch30

Foygel, R., & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In *Advances in neural information processing systems* (pp. 604–612). Retrieved from http://arxiv.org/abs/1011.6640%5Cnhttp://papers.nips.cc/paper/4087-extended-bayesian-information-criteria-for-gaussian-graphical-models doi: 10.1.1.231

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10. doi: 10.1007/s00127-016-1319-z

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. doi: 10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–24. Retrieved from http://www.jstatsoft.org/v33/i01/ doi: 10.18637/jss.v033.i01

Friedman, J., Hastie, T., & Tibshirani, R. (2018). *Package 'glasso'*. Retrieved from http://statweb.stanford.edu/~tibs/glasso/

Fu, W., & Knight, K. (2000, 10). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378. Retrieved from http://projecteuclid.org/euclid.aos/1015957397 doi: 10.1214/aos/1015957397

Gauraha, N., & Swapan, P. (2018). Constraints and Conditions: the Lasso Oracle-inequalities. *arXiv*. Retrieved from http://arxiv.org/abs/1603.06177

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems*, *96*(1), 86–103. doi: 10.1016/j.biosystems.2008.12.004

Homrighausen, D., & McDonald, D. J. (2013). Risk-consistency of cross-validation with lasso-type procedures. (2), 1–25. Retrieved from http://arxiv.org/abs/1308.0810

Homrighausen, D., & McDonald, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning*, *97*(1-2), 65–78. doi: 10.1007/s10994-014-5438-z

Hsu, C.-W., Sinay, M. S., & Hsu, J. S. J. (2012, 4). Bayesian estimation of a covariance matrix with flexible prior specification. *Annals of the Institute of Statistical Mathematics*, *64*(2), 319–342. Retrieved from http://link.springer.com/10.1007/s10463-010-0314-5 doi: 10.1007/s10463-010-0314-5

Kossakowski, J. J., Epskamp, S., Kieffer, J. M., van Borkulo, C. D., Rhemtulla, M., & Borsboom, D. (2016). The application of a network approach to Health-Related Quality of Life (HRQoL): introducing a new method for assessing HRQoL in healthy adults and cancer patients. *Quality of Life Research*, *25*(4), 781–792. doi: 10.1007/s11136-015-1127-z

Kuismin, M., & Sillanpää, M. J. (2016, 2). Use of Wishart Prior and Simple Extensions for Sparse Precision Matrix Estimation. *PLOS ONE*, *11*(2), e0148171. Retrieved from http://dx.plos.org/10.1371/journal.pone.0148171 doi: 10.1371/journal.pone.0148171

Kuismin, M. O., & Sillanpää, M. J. (2017, 11). Estimation of covariance and precision matrix, network structure, and a view toward systems biology. *Wiley Interdisciplinary Reviews: Computational Statistics*, *9*(6), e1415. Retrieved from http://doi.wiley.com/10.1002/wics.1415 doi: 10.1002/wics.1415

Lauritzen, S. L. (1996). *Graphical models (Vol. 17)*. Clarendon Press.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability Approach

to Regularization Selection (StARS) for High Dimensional Graphical Models. *Advances in neural information processing systems*, 1432–1440. 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/1306.0976http://projecteuclid.org/euclid.aos/1388545674 doi: 10.1214/13-AOS1169

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0005796716301103 doi: 10.1016/j.brat.2016.06.006

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

Nichols, T., & Holmes, A. (2003). Nonparametric Permutation Tests for Functional Neuroimaging. *Human Brain Function: Second Edition*, *25*(August 1999), 887–910. doi: 10.1016/B978-012264841-0/50048-2

Pereira-Morales, A. J., Adan, A., & Forero, D. A. (2017). *Network analysis of multiple risk factors for mental health in young Colombian adults* (Vol. 0) (No. 0). Taylor & Francis. Retrieved from https://doi.org/10.1080/09638237.2017.1417568 doi: 10.1080/09638237.2017.1417568

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1), 43–53. Retrieved from http://dx.doi.org/10.1016/j.newideapsych.2011.02.007 doi: 10.1016/j.newideapsych.2011.02.007

Tibshirani, R. (2015). Sparsity and the Lasso. In *Statistical machine learning* (pp. 1–15).

Williams, D. R., & Rast, P. (2018). Back to the basics: Rethinking partial correlation network methodology. *PsyArXiv*, 1–15. doi: 10.17605/OSF.IO/FNDRU

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2018). On Non-Regularized Estimation of Psychological Networks. *PsyArXiv*.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. Retrieved from http://journals.sagepub.com/doi/10.1177/1745691617693393 doi: 10.1177/1745691617693393

Yu, Y., & Feng, Y. (2014). Modified Cross-Validation for Penalized High-Dimensional Linear Regression Models. *Journal of Computational and Graphical Statistics*, *23*(4), 1009–1027. doi: 10.1080/10618600.2013.849200

Zhang, P. (1993). Model Selection Via Multifold Cross Validation. *The Annals of Statistics*, 299–313. Retrieved from https://www.jstor.org/stable/3035592

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, *7*, 2541–2563. Retrieved from http://dl.acm.org/citation.cfm?id=1248637 doi: 10.1109/TIT.2006.883611

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge Package for High-dimensional Undirected Graph Estimation in R. *Journal of machine learning research : JMLR*, *13*, 1059–1062. Retrieved from http://dl.acm.org/citation.cfm?id=2343681%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26834510%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4729207 doi: 10.1002/aur.1474.Replication

Zhu, Y., & Cribben, I. (2018). Graphical models for functional connectivity networks: best methods and the autocorrelation issue. *Brain Connectivity*, *8*(3), 139–165. Retrieved from http://www.liebertpub.com/doi/10.1089/brain.2017.0511 doi: 10.1089/brain.2017.0511