MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 5, 990–1012

OPEN ACCESS

Missing Data Handling via EM and Multiple Imputation in Network Analysis using Glasso and Atan Regularization

Kai Jannik Nehler and Martin Schultze

Department of Psychology, Goethe University Frankfurt, Frankfurt am Main, Germany

ABSTRACT

The existing literature on missing data handling in psychological network analysis using cross-sectional data is currently limited to likelihood based approaches. In addition, there is a focus on convex regularization, with the missing handling implemented using different calculations in model selection across various packages. Our work aims to contribute to the literature by implementing a missing data handling approach based on multiple imputation, specifically stacking the imputations, and evaluating it against direct and two-step EM methods. Standardized model selection across the multiple imputation and EM methods is ensured, and the comparative evaluation between the missing handling methods is performed separately for convex regularization (glasso) and nonconvex regularization (atan). Simulated conditions vary network size, number of observations, and amount of missingness. Evaluation criteria encompass edge set recovery, partial correlation bias, and correlation of network statistics. Overall, missing data handling approaches exhibit similar performance under many conditions. Using glasso with EBIC model selection, the two-step EM method performs best overall, closely followed by stacked multiple imputation. For atan regularization using BIC model selection, stacked multiple imputation proves most consistent across all conditions and evaluation criteria.

KEYWORDS
Network analysis; missing values; simulation study; regularization; multiple imputation

Network analysis to model psychological constructs or disorders, termed as *psychological networks*, is emerging as a popular tool, particularly for clinical research (e.g., Fried et al., 2015; Lu et al., 2023; Miers et al., 2020), but has also in areas such as personality research (e.g., Jefferies et al., 2023) or health psychology (e.g., van Zyl, 2021). The methodology can be utilized in both longitudinal and cross-sectional settings, with the present paper focusing on the latter. While there is extensive debate regarding the best approach to model estimation and selection through theoretical considerations or simulation studies (e.g., Isvoranu & Epskamp, 2023; Williams & Rast, 2020), the understanding of the impact of missing values and potential ways to handle them remains limited. This issue is further highlighted by recently published reporting standards (Burger et al., 2023), which only briefly touch upon the treatment of missing values.

However, it is well established that missing values can lead to errors in the analysis, especially if not handled correctly (Schafer & Graham, 2002). Edge weights connected to nodes with missing data may be distorted, potentially affecting whether those edges are included in the final model. Consequently, there is a pressing need to investigate missingness in network estimation. Mansueto et al. (2023) found two different missing data handling approaches (full information maximum likelihood estimation and Kalman filter imputation) to perform quite well when faced with simulated data missing completely at random in longitudinal settings. Current knowledge on handling missing data in cross-sectional psychological network analysis (e.g., Falk & Starr, 2023; Nehler & Schultze, 2024b) has primarily focused on the performance of approaches based on adaptations of the expectation-maximization (EM) algorithm (originally introduced by Dempster et al., 1977) and has only investigated a single method for network estimation, namely the graphical lasso (glasso; Friedman et al., 2008). To broaden the toolset for handling missing data in cross-sectional network analysis, we introduce a multiple imputation approach and investigate its performance, as well as that of the current EM-based methods in two distinct network analysis techniques.

Network estimation and model selection

In psychological networks, variables (e.g., questionnaire items or symptoms of a disorder) are represented by nodes and their relationships by edges. For continuous variables, edges often reflect partial correlations, capturing the unique association between two variables. A partial correlation ρ for two variables j and j' (with $j \neq j'$) can be modeled by using the inverse covariance matrix $\Theta = \Sigma^{-1}$, which we will refer to as the *precision matrix* (see Lauritzen, 1996).

$$\rho_{jj'} = -\frac{\Theta_{jj'}}{\sqrt{\Theta_{jj}}\sqrt{\Theta_{j'j'}}} \tag{1}$$

The proportion of edge weights in a network that are not equal to zero is called its *density*, with highly connected networks called *dense*, while less connected ones are called *sparse*. Cross-sectional networks do not distinguish between intra- and inter-individual variation (Hamaker, 2012) and present the edges as undirected (Epskamp & Fried, 2018). In practice, the calculation in Equation (1) must be performed based on the estimated precision matrix $\hat{\Theta}$, since the true matrix is unknown. During the estimation, sampling variation can lead to spurious edges (Costantini et al., 2015). Consequently, various methods for model estimation and selection have been proposed—see Isvoranu and Epskamp (2023) for a comprehensive but not exhaustive overview. In addition to glasso, which has been examined in the context of missing data (e.g., Falk & Starr, 2023; Nehler & Schultze, 2024b), we also investigate a nonconvex alternative in this study, specifically focusing on the atan penalty (Williams, 2020).

For the scope of this paper, we limit ourselves to the case of continuous, multivariate normally distributed variables. Models with this type of data are termed *Gaussian graphical models* (GGM; Costantini et al., 2015; Lauritzen, 1996). We denote the observed, centered variables as Y following previous research (Epskamp & Fried, 2018; Williams & Rast, 2020). The sample covariance matrix is denoted as S, and the population covariance matrix as Σ.

Convex regularization

The glasso (also called ℓ_1 regularization) estimates the precision matrix that maximizes a penalized log-likelihood based on the sample covariance matrix.

$$\log\ \det(\Theta)\text{-tr}(S\Theta) - \lambda||\Theta||_1 \tag{2}$$

The last part of the equation is a convex penalty term, which uses the penalty parameter λ to scale the sum of all off-diagonal absolute values of the precision matrix Θ. Thus, maximizing this equation results in shrinking entries in the precision matrix—in some cases to 0. Regularization is performed with varying penalty parameters λ logarithmically spaced between a minimum and a maximum value (Epskamp, 2016). Commonly, 100 values for λ are used, returning as many estimated precision matrices $\hat{\Theta}_a$ with $a \in \{1, \ldots, 100\}$.

Following the estimation process, various methods can be employed to determine the optimal one among the resulting $\hat{\Theta}_a$. Most of these methods are based on unpenalized log-likelihoods, which are also computed using the sample covariance matrix S.

$$ll(\hat{\Theta}_a) = \frac{n}{2}(\log\ \det(\hat{\Theta}_a) - \text{tr}(S\hat{\Theta}_a) - p \cdot \log(2\pi)) \tag{3}$$

Two crucial considerations arise: First, the presented unpenalized log-likelihood is always best for the least penalized $\hat{\Theta}_a$. Second, as the number of observations increases, the disparity between the log-likelihoods of sparse and dense networks grows. Thus, instead of performing model selection on the unpenalized log-likelihood, the *Extended Bayesian Information Criterion* (EBIC; Foygel & Drton, 2010) is used, aiming to identify the model with the lowest value among the 100 candidates:

$$EBIC_a = -2ll(\hat{\Theta}_a) + |E_a|\log n + 4|E_a|\gamma \log p \tag{4}$$

E_a represents the edge set of the respective graph with $|E_a|$ being the number of non-zero elements in the upper triangle of the corresponding estimated precision matrix $\hat{\Theta}_a$. The inclusion of this term serves to counterbalance the influence of the log-likelihood by imposing a higher penalty on denser networks. The hyperparameter γ serves as an additional penalty for complex models. Simulations indicated an optimal value of $\gamma = 0.5$ (Foygel & Drton, 2010), outperforming higher penalizing values when using a network structure resembling the expected structure of psychological constructs as the population (Epskamp, 2016). In the following, we adhere to the nomenclature proposed by Williams et al. (2019) and refer to the combination of glasso regularization to estimate a model and EBIC to achieve selection as $glasso_{EBIC}$.

In cases with a similar number of observations and nodes, the performance of $glasso_{EBIC}$ stands out due to its specificity, where the edges included in a chosen network structure can be confidently regarded as genuine (Epskamp, 2016). It returns mostly sparse network structures, developed to provide applied researchers with a well-interpretable network structure. However, Williams et al. (2019) show that the sensitivity of the approach varies depending on the number of observations, returning more non-zero edges as n increases. In their simulation study the true model is not found for $n \to \infty$ if the true network structure is not extremely sparse, which is uncommon in psychological practice (Wysocki & Rhemtulla, 2021). One explanation is that the $\beta_{\min}$ condition required for consistency may not be satisfied, meaning that the smallest entries in the inverse covariance matrix are not large enough to survive regularization (Zhao & Yu, 2006). However, such small entries—resulting from weak partial correlations—are a common feature of psychological networks (Wysocki & Rhemtulla, 2021). Another contributing factor is that, as n increases, sensitivity improves while specificity declines, resulting in denser networks that deviate further from the true structure. It is important to note that this effect is specific to the standard *glasso*$_{EBIC}$ implementation in the well-known *qgraph* package (Epskamp et al., 2012) and similar implementations, which do not include an explicit mechanism for controlling the false positive rate. A further limitation of the standard *glasso*$_{EBIC}$ implementation is that all edges are penalized, including those that are ultimately retained in the final network structure (Williams & Rast, 2020).

Nonconvex regularization

Nonconvex regularization was introduced to psychological network analysis literature by Williams (2020) with the objective of mimicking a best subset selection approach, while maintaining computational efficiency by avoiding the need to test every conceivable edge combination. The underlying assumption is that universal consistency can be attained through minimal shrinkage on big parameters, while effectively shrinking small parameters to zero (Zhao & Yu, 2006). Williams (2020) demonstrated nonconvex regularization approaches to meet these criteria for network analysis, with the atan penalty Wang and Zhu (2016) showing the most promising performance. However, a known limitation of nonconvex penalties is the lack of a guaranteed unique global optimum (Williams, 2020).

In general, like the *glasso*$_{EBIC}$, nonconvex regularization operates by estimating the precision matrix through maximizing a penalized log-likelihood. Yet, it offers greater flexibility in the penalty term through individualized calculation depending on each entry $\Theta_{jj'}$.

$$\log\ \det(\Theta) - \mathrm{tr}(S\Theta) - \sum_{j \neq j'} q_{\lambda,\eta}\left(|\Theta_{jj'}|\right) \qquad (5)$$

In the case of atan regularization, the individual contribution to the penalty term of each entry θ in the precision matrix is computed as follows:

$$q_{\lambda,\eta}(\theta) = \lambda\left(\eta + \frac{2}{\pi}\right)\arctan\left(\frac{|\theta|}{\eta}\right) \qquad (6)$$

λ and η (with $\eta > 0$) represent the tuning parameters for the penalty function. As was the case for glasso, the diagonal of Θ is excluded from the penalization. As η approaches infinity, this results in ℓ_1 regularization, while η tending toward zero approximates best subset selection. Sparsity is affected by λ, where $\lambda \to 0$ results in the maximum likelihood estimation of Θ, which represents a non-regularized structure. $\lambda \to \infty$ results in stronger regularization and ultimately an empty network. Notably, Williams (2020) demonstrated that the decision regarding which parameter to fix and which to vary does not have a significant impact and freely selecting both parameters does not lead to an improvement in performance. In our simulation, we varied the parameter λ, while keeping η fixed at a value of 0.01. Typically, 50 different values are considered for λ logarithmically spaced between a lower and an upper bound. This results in estimating 50 precision matrices $\hat{\Theta}_b$ with $b \in \{1, ..., 50\}$.

As previously stated, a critique of nonconvex penalties is the absence of a guaranteed global optimum (Williams, 2020). Fan et al. (2014) argued that approximating the optimal solution is acceptable, provided that the results maintain desired properties such as consistency in model selection. Under the assumption that the number of non-zero elements is known, they demonstrated this to be true for precision matrices. Nevertheless, such an approximation still does not guarantee a global optimum for every single application case. There are several different possible algorithms for the approximation. We have chosen to use the one-step estimator (Zou & Li, 2008), because of its computational efficiency and performance in cases with $n \gg p$. This condition should be sufficiently met in psychological settings, typically with a higher number of observations compared to the amount of variables (Wysocki & Rhemtulla, 2021).

Selecting the final model among the 50 $\hat{\Theta}_b$ is accomplished using information criteria. In their simulation, Williams (2020) employed the *Bayesian Information Criterion* (BIC; originally introduced by Schwarz, 1978), which corresponds to Equation (4) with $\gamma = 0$. The author showed that combining the atan penalty with BIC for model selection resulted in increased sensitivity as sample size grew. Additionally, specificity remained high for small n and was not impacted strongly with an increasing number of observations or varying sparsitiy in the population network. Henceforth, we refer to the combination of atan regularization with BIC model selection as $atan_{BIC}$.

The described methods of estimating network structures operate with completely observed data sets and their performance in these situations has been compared elsewhere (e.g., Isvoranu & Epskamp, 2023; Williams, 2020). In this study, we intend to evaluate their respective performance when used in conjunction with one MI approach, as well as two EM based approaches, in the presence of missing values. It is important to note that it is not the intention of this study to compare the performance of $glasso_{EBIC}$ and $atan_{BIC}$ directly. Following reviews, however, we provide tentative comparisons in the electronic supplemental material (ESM).

Missing values and handling approaches

Traditionally, the occurrence of missing values is differentiated into three different missing mechanisms: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR) (Little & Rubin, 2019; Rubin, 1976). With data MCAR, the missingness is independent of any properties of the observed or missing data. If the data is MAR, missingness depends only on observed values. Under MNAR, the probability of missingness is additionally conditional on unobserved values. In the following paragraphs, we review EM-based approaches for handling missing data in network estimation, introduce our proposed use of multiple imputation in this context, and discuss potential differences in performance between these methods.

Two-step EM

The two-step EM approach handles missing data by applying the EM algorithm in the first step, followed by model estimation and selection in the second step. During the *expectation-step* (E-step) of the EM algorithm, missing entries in the data matrix are filled with conditional expectations. These conditional expectations are subsequently employed to compute sufficient statistics, with a residual term added if both variables were missing for an observation (Little & Rubin, 2019). In the following *maximization-step* (M-step), (co-)variances are computed using the sufficient statistics. The E- and the M-step are performed iteratively until convergence is achieved. It is worth noting that the use of *full information maximum likelihood* (FIML) yields the same results as the EM algorithm used here when estimating unconstrained covariance matrices (Enders, 2001), as is done in the first step of this approach. In the second step, network estimation and model selection are performed based on the estimated covariance matrix.

To our knowledge, the two-step EM is currently only evaluated for $glasso_{EBIC}$ but can easily be extended to $atan_{BIC}$. The procedure is implemented in the R package *bootnet* (Epskamp et al., 2018), which offers extended options for network analysis on top of the *qgraph* package. As described, the packages uses the covariance matrix generated by the EM algorithm to perform model estimation. For model selection using information criteria, as per Equation (4), as well as Equation (3), the covariance matrix from the EM algorithm also serves as the sample covariance matrix S, while several methods for determining the sample size in the presence of missing values are provided.

In a recent evaluation conducted by Nehler and Schultze (2024b), the authors employed the default settings of *qgraph* for $glasso_{EBIC}$, wherein the sample size for log-likelihood and EBIC calculation is set to the average of all pairwise sample sizes for variances and covariances. This evaluation showed that the method yields highly sparse networks under various conditions, with only a limited number of true edges being detected, especially when dealing with increased missing data and a small number of observations. Falk and Starr (2023) evaluated the two-step EM using the average pairwise sample size of only covariances instead with similar results.

Direct EM

Städler and Bühlmann (2012) proposed an integrated EM algorithm as a combination of missing data handling and glasso regularization of the graph structure. This method shares the same E-step with a standard EM algorithm, computing conditional expectations, residual terms and sufficient statistics. In the M-step, the algorithm computes the inverse covariance matrix based on the sufficient statistics and subsequently employs regularization to produce a sparse precision matrix. This resulting matrix is then utilized to compute conditional expectations in the next E-step. The two steps are repeated until convergence is achieved. The algorithm is run separately for different penalty parameters λ, typically resulting in 100 estimated precision matrices $\hat{\Theta}_a$. Since this approach does not involve a separation between model estimation *via* glasso regularization and missing data handling, we refer to it as direct EM.

An evaluation by Städler and Bühlmann (2012) demonstrated promising results with sparse matrices—representing typical network structures in the field of biology—showcasing sensitivity and specificity at the same time. While the authors included model selection *via* EBIC in their simulations, they did not describe the specific technique used to calculate the information criteria, which has led to varying implementations in practice. The R package *cglasso* by Augugliaro et al. (2023) employs the EBIC, as defined previously in Equation (4). It utilizes the total number of observations (with or without missingness). For S, it uses the inverse of the non-penalized precision matrix (i.e., the assumed sample covariance matrix without regularization) in the final iteration of the M-step. Therefore, each $\hat{\Theta}_a$ matrix is compared to a distinct sample covariance matrix, taken from the respective estimation process. Nehler and Schultze (2024b) investigated the direct EM using the *cglasso* package with population networks estimated from psychological data, which are typically denser than those investigated by Städler and Bühlmann (2012). The results indicated high sensitivity at the potential risk of losing specificity.

To our knowledge the direct EM approach has not been extended to include $atan_{BIC}$ until now. Our implementation is based on the $glasso_{EBIC}$ procedure Städler and Bühlmann (2012) with only minor modifications, which are describe in detail in Appendix B. The alteration involves using the atan penalty for regularization in the M-step, and the best model is selected among competing models *via* BIC to align with the two-step and complete data approaches. The calculation of the information criteria to perform model selection is described later in the section *Model selection with missing values*.

Multiple imputation

The fundamental concept of multiple imputation (MI) involves replacing each missing data point with several independent values, thereby creating several complete versions of the data set (Rubin, 1987a, 1996). We used *predictive mean matching* (PMM) as the imputation method, which has been shown to be effective across various scenarios, although its performance may falter with smaller n (Kleinke, 2017). PMM operates by first predicting values for the variable that is imputed for all individuals. The prediction incorporates noise and parameter uncertainty by drawing parameters from their posterior distribution (van Buuren, 2018). PMM then identifies observed cases with predicted values similar to those of the missing case and randomly draws one as the donor for imputation.

In multivariate missing data settings, *fully conditional specification* (FCS) is often used (van Buuren, 2006, 2007). This approach specifies conditional distributions for each variable separately, eliminating the need for a joint multivariate model. Imputations across variables are generated iteratively—while five iterations may be sufficient for unbiased estimates, higher correlations or increased missingness may require additional iterations (van Buuren et al., 1999). Regarding the number of imputations, a higher number improves replicability, but as few as two imputations may suffice for generating point estimates (von Hippel, 2020). Since different imputations yield different imputed values, *Rubin's rules* are applied to pool resulting statistical parameters, including the computation of their standard errors (Rubin, 1987b).

Using the traditional approach to pool results of multiply imputed data, selecting a final network structure is challenging because edges may be included in some imputations but excluded in others. Pooling raw values, while treating non-existent edges as zero in mean calculations, could introduce bias. We propose an approach inspired by variable selection in multiple regression (Wood et al., 2008), whereby the multiply imputed data sets are stacked into one and model estimation is performed only once on the combined data set—an approach that has also been applied in the context of *structural equation modeling* (SEM; Lang & Little, 2014). Evaluation of the stacking approach indicated a need for standard error corrections but unbiased point estimates (van Buuren, 2018; Wan et al., 2015), making it suitable for exploratory network analysis *via* $atan_{BIC}$ or $glasso_{EBIC}$.

In summary, our proposal involves imputing data sets using PMM and FCS, stacking them into a single data set, and deriving the corresponding covariance matrix, as also done by Lang and Little (2014). Subsequently, regularization is applied using either glasso or atan to obtain estimates for $\hat{\Theta}_a$ or $\hat{\Theta}_b$. Details on model selection are provided in the corresponding section.

Comparison of missing data handling approaches

Direct and two-step EM approaches haven been evaluated in both the SEM and network analysis literature. In SEM, Savalei and Bentler (2009) and Zhang and Savalei (2023) compared full information maximum likelihood (FIML)—a method conceptually similar to the direct EM approach described above—with two-step procedures. Their findings suggest that while two-step methods yield valid and consistent estimates, they tend to be less efficient. Falk and Starr (2023) argue that this reduced efficiency may result from a loss of information about uncertainty in specific covariance elements when transitioning from a saturated to a structured model in the second step. In the context of network analysis, Nehler and Schultze (2024b) found that the two-step EM approach performed considerably worse than direct EM, particularly in terms of sensitivity. This pattern reversed only under conditions of high missingness and large sample sizes. Similarly, Falk and Starr (2023) reported inferior performance of the two-step EM in recovering population networks, although the observed differences were smaller. Notably, the procedures used for model selection varied not only between these two studies, but also within each study depending on the missing data handling approach. These methodological differences are discussed in the following section.

While likelihood based methods are often considered the gold standard for handling missing data (van Buuren, 2018), MI offers greater flexibility, particularly when dealing with non-normal data and nonlinear relationships in the imputation model (Vink & Van Buuren, 2013). Since MI has not yet been evaluated in the context of network analysis, relevant insights regarding the comparison of missing data handling methods must be drawn from other domains. Within donor-based imputation approaches, Jia and Wu (2023) identified PMM as particularly effective under non-normality in SEM contexts. Moreover, Lee and Shi (2021) demonstrated that MI estimates remained stable under model misfit, whereas FIML-based approaches tended to deviate more from full-data results as the degree of misfit increased. This is particularly relevant for the present study, as regularization in network models introduces intentional model misfit, which may disproportionately affect the performance of direct EM.

Lang and Little (2014) investigated the use of stacked data sets to compute a single covariance matrix (a *supermatrix*) in SEM. However, their evaluation focused exclusively on significance testing and convergence, with the stacking procedure demonstrating advantages over FIML in terms of convergence. In principle, the stacked MI approach resembles the approach utilized in the two-step EM method in the sense that missing data handling is used to estimate the covariance matrix in the first step and regularization is applied afterwards. As such, if all distributional assumptions hold, both approaches are asymptotically equivalent (Lee & Shi, 2021; Pigott, 2001). Nevertheless, this does not necessarily mean they will yield identical results in practice, as highlighted by Pigott (2001) in the context of regression.

The literature suggests that all three investigated approaches should exhibit similar consistency. In terms of efficiency, the direct EM approach may be superior because it retains information about uncertainty in the estimated parameters. However, the model misfit introduced by regularization may disproportionately affect the consistency and efficiency of the direct EM approach. Since this misfit stems from the regularization procedure, its impact may differ between the $glasso_{EBIC}$ and $atan_{BIC}$. Drawing expectations from prior studies in network analysis is challenging because those studies vary substantially in how information criteria are computed, whereas the approach we chose in this study follows a standardized procedure.

Model selection with missing values

As outlined earlier, the introduction of missing values necessitates specific choices in the calculation of the log-likelihood and subsequently in the evaluation criteria. Existing implementations and previous studies have varied in their approaches. For instance, Augugliaro et al. (2023) selected the estimate from the final EM algorithm iteration as S for Equation (3) in their implementation of the direct EM. However, these estimates vary across penalty parameters, complicating the claim that they truly represent the sample covariance matrix. In contrast, the two-step EM and stacked MI approaches could simply use the estimated covariance matrix from the first step.

Furthermore, as previously stated, the determination of the number of observations for multiplication in Equation (3) has strong influence on the following model selection. Sparse models among the candidate solutions typically exhibit worse fit when using unpenalized log-likelihood, a discrepancy that becomes more pronounced as n increases. Thus, a larger n introduces greater variation in log-likelihood values used for information criteria calculation, which tends to favor denser networks among the candidates. In contrast, the impact of a higher n on the incorporated log-term is relatively minor. The two-step implementation (Epskamp et al., 2018) employs pairwise averaging to calculate n, whereas the direct EM method by Augugliaro et al. (2023) uses the total observation count. This discrepancy may explain the findings of Nehler and Schultze (2024b), who observed that the two-step approach produced more sparse and specific results, whereas the direct EM method yielded networks with greater density and sensitivity.

A similar pattern regarding the calculation of log-likelihood and its impact on the comparison of missing data handling approaches can be assumed for the study of Falk and Starr (2023). In their study, the two-step EM approach uses the pairwise average number of observations for the covariances in calculating the log-likelihood with Equation (3), sometimes resulting in a very small effective sample size. This can lead to the selection of sparser networks among the candidate models. In contrast, the direct EM method, which employs the observed data log-likelihood (as described below), generates larger discrepancies among candidate models. In their simulation, the direct EM method with the observed data log-likelihood was found to be more sensitive, while the two-step EM method, relying on the pairwise number of observation and the matrix-based calculation of the log-likelihood, tended to produce sparser networks. These differences between the missing data handling approaches could potentially be attributed to variations in the log-likelihood and information criteria methods.

To emphasize the specific evaluation of the missing handling we integrated a consistent approach in computing the information criteria across all three missing data handling methods. In accordance with Falk and Starr (2023), we use the observed data log-likelihood instead of having to choose an n and the sample covariance matrix. The observed data log-likelihood is computed as the sum of the individual log-likelihoods for each observation i:

$$ll(\hat{\Theta},\hat{\mu}) = \sum_{i=1}^{n}\left(-\frac{p_i}{2}\cdot\log(2\pi) - \frac{1}{2}\log\left|\left(\hat{\Theta}^{-1}\right)_i\right| - \frac{1}{2}(Y_i-\hat{\mu}_i)^T\hat{\Theta}_i(Y_i-\hat{\mu}_i)\right) \tag{7}$$

Here, Y_i represents the observed centered variables for the specific individual, and all other values in the calculation (model parameters $\hat{\Theta}$, p and means $\hat{\mu}$) are reduced to those that include the observed variables of that specific individual. Centering the observed variables with missing values based on sample information does not imply that the true mean of the centered variable is zero. Therefore, the mean is also incorporated into the equation and estimated based on the stacked data matrix in the stacked MI approach, *via* the EM algorithm in the first step of the two-step EM approach, and naturally returned by the direct EM along with the estimate for $\hat{\Theta}$. Notably, the resulting log-likelihood value without missing data would be the same as Equation (3).

Although the contribution of the term $\log(n)$ to the information criteria calculation may be relatively minor, it nevertheless requires specification. This term should represent the available information. We elected to utilize the average pairwise observations of all covariances, motivated by the rationale that the diagonal elements (variances) are not subject to regularization.

The present study

As outlined at the beginning of this manuscript, the present study aims to evaluate a stacked MI approach against EM algorithms under consistent model estimation and selection criteria. We integrated all missing data handling techniques with both $glasso_{EBIC}$ and $atan_{BIC}$. Given that all three missing data handling methods are modern and conceptually similar—particularly stacked MI and two-step EM—we expect them to perform comparably. However, evidence from previous literature suggest that differences in consistency and efficiency may arise across approaches, although the direction of these effects remains unclear.

Methods

A Monte Carlo simulation study was conducted to compare the performance of the missing data handling techniques in estimating networks using both $glasso_{EBIC}$ and $atan_{BIC}$. We used an openly available data set containing responses to the Fisher Temperament Inventory (Brown et al., 2013; Fisher et al., 2010)[1]to construct the population networks from which the data were simulated. This choice aligns with the approach taken in other simulation studies investigating psychological network analysis (e.g., Isvoranu & Epskamp, 2023; Mansueto et al., 2023), aiming to mimic a realistic psychological scenario.

A similar density value of approximately 0.32 was maintained across all population networks, which falls within the range of sparse structures typically encountered in applied psychological research (Wysocki & Rhemtulla, 2021). Population network structures were achieved by removing edges with the lowest partial correlations until the specified density was reached, while ensuring that there were no isolated nodes. After identifying the edges to be removed, the precision matrix was re-estimated, i.e., the identified edges were forced to zero without applying any regularization to the remaining edges. In line with the work of Wysocki and Rhemtulla (2021), we did not ensure that the assumptions required for the consistency of

[1] Data can be retrieved from http://openpsychometrics.org/FTI_data.zip/.

model selection were met by the population networks, instead focusing the partial correlations on the realistic psychological data set. More detailed information on the population networks—specifically, the distribution of partial correlations and the detectability of edges—can be found in Appendix A.

Simulation design

In the generation of the data set, three factors were manipulated. Network size (p) was varied among 8, 24, and 48. The choices of 8 and 24 fall within the typical range of network sizes encountered in psychological research (Wysocki & Rhemtulla, 2021), while 48 represents a more extreme scenario for future reference. Sample size (n) was adjusted across three levels: 400, 800, and 1600 Finally, the rate of missingness (m) was varied among completely observed, 0.1, 0.2, and 0.3. Nehler and Schultze (2024b) explored different missing data mechanisms but found minimal distinctions between MAR and MNAR, attributing this surprising result to the strong interdependence of nodes within the network. Based on these findings, we elected to focus exclusively on MAR. The data generation conditions were crossed with the three missing data handling methods, excluding any redundant combinations arising from the intersection of these factors, resulting in a total of 90 unique conditions. Each condition was replicated 500 times, and in cases of missingness, all three handling methods were tested. The simulation was conducted once for $atan_{BIC}$ and once for $glasso_{EBIC}$.

Data creation and missing values generation

Data were simulated from the multivariate normal distribution with the correlation matrix derived from the population networks described above. This initially resulted in completely observed data sets. Replications were simulated in a way that conditions based on the same parameters in data generation (network size and number of observations) returned the same 500 data sets. For conditions with missing values, observations were removed from the complete data sets. This approach not only facilitates comparability among different missing data handling methods but also allows for a direct comparison between a given condition using complete and incomplete data sets.

Missing values generation was done with a slight modification of the approach proposed by Grund et al. (2018). Let R be a matrix including the latent response propensities.

$$R_{ij} = \beta_{1jj'} \cdot P_{ij'} + r_{ij} \tag{8}$$

The response propensity of observation i on variable j is denoted with R_{ij}. $\beta_{1jj'}$ is the regression parameter for the missing propensity of a variable j with a standardized, completely observed predictor $P_{j'}$ (with $\beta_{1jj'}$ set to 0.7). r_{ij} represents a normally distributed residual value with a mean of zero and a variance of $1 - \beta^2_{1jj'}$. Critical values R_c were determined according to the missing proportion of the specific conditions. Any data point with $|R_{ij}| > |R_c|$ was deleted from the complete data set. Missing values generation was aimed at representing MAR ($j \neq j'$). In all conditions with $m > 0$, half of the items contained missing values, while the other half were completely observed. Thus, in conditions with $m = 0.1$, variables j were 20% missing. Variables with and without missingness were determined to have a similar average in the network statistic strength.

Evaluation criteria

Simulation studies evaluating network analysis tools provide a wide range of possible criteria. In our manuscript, we include evaluation criteria offering distinct perspectives: encompassing the examination of the edge set, parameter estimation, and recovery of network statistics[2].

Edge set

The first evaluation criteria focus on discerning the network structure. Network density is defined as the ratio of non-zero estimated edge weights to the total number of possible edges. Beyond determining the correct density, precision in edge selection is crucial and is commonly assessed using sensitivity and specificity (e.g., Isvoranu & Epskamp, 2023; Williams, 2020). Specificity captures the accuracy in identifying edges with weights of zero, while sensitivity quantifies the accuracy in identifying non-zero weighted edges.

For all handling mechanisms, an increase in missingness is expected to result in a lower density. With $glasso_{EBIC}$, the densities of the estimated networks are projected to exceed the population density for larger sample sizes. This would likely be accompanied by a decrease in specificity and an increase in sensitivity. Conversely, when combining missing handling mechanisms with $atan_{BIC}$, the estimated densities should approach the population density with increasing sample size, consistently maintaining

[2] Notable evaluation criteria not discussed include for example the Matthew correlation coefficient (Matthews, 1975) or the Kullback-Leibler loss (Kullback & Leibler, 1951). They can be examined using the evaluation code in the corresponding repository on the Open Science Framework (https://osf.io/kdv7t/).

heightened specificity while experiencing a smaller increase in sensitivity.

Parameter estimation

In network analysis, the primary parameters of interest are the edge weights, denoted by partial correlations. Bias can occur when non-zero edge weights are inaccurately estimated or when genuine zero edge weights are erroneously omitted. These potential biases are reflected in the previously mentioned evaluation criteria of sensitivity and specificity. To complement the bias introduced by edge set identification, here, we calculate the difference between estimated and true partial correlations for correctly identified non-zero edges. To derive more comprehensible aggregate performance metrics, we calculated the average raw bias for each replication.

For the $glasso_{EBIC}$ approach, we anticipate a negative bias that intensifies with greater levels of missing data. For the $atan_{BIC}$ method, we do not have a predefined expectation regarding the direction of the bias. However, we do expect performance to degrade as the proportion of missing data increases. For increases in sample size, we expect this bias to become less pronounced (for $glasso_{EBIC}$) or approach zero (in the case of $atan_{BIC}$).

Network statistics

An important consideration for applied researchers is the use of descriptive statistics for the resulting network, such as centrality indices (Epskamp et al., 2018; Opsahl et al., 2010), which are widely applied to identify the most influential nodes (e.g., Lu et al., 2023). However, among these metrics, only strength emerges as a reliable and theoretically sound measure (e.g., Bringmann et al., 2019; Epskamp et al., 2018). Given this, we only evaluate estimated strength values in our simulation. Following the conceptualization by Barrat et al. (2004), the strength of a node in weighted networks is defined as the sum of the absolute weight values of its edges. Aligning with the methodology of Epskamp and Fried (2018), we utilized the correlation of strength values between the population and the estimated networks as an evaluation criterion.

We anticipate similar behavior of the performance criteria across conditions for the analyses using $glasso_{EBIC}$ and $atan_{BIC}$, although predictions are less definitive compared to other evaluation criteria. Specifically, we expect correlations to decrease as the proportion of missing data increases across all missing data handling techniques. However, as the ratio n/p increases, we believe that the correlations will more closely resemble those derived from complete data.

Software and parameter settings

The simulations and evaluations were conducted using R (Core Team, 2023). The code for these processes, including our customized version of the direct EM and the computation of individual log-likelihoods, can be accessed on the Open Science Framework (https://osf.io/kdv7t/). A more detailed description of our implementation of the direct EM can also be found in Appendix B.

Observations from the population networks were drawn using the *MASS* package (Venables & Ripley, 2002, Version 7.3.60). Notably, all variables were standardized in our simulation. For the two-step approach, correlation matrices were estimated using the EM algorithm by the *lavaan* package (Rosseel, 2012, Version 0.6.16). MI was conducted using the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011, Version 3.16.0) with 20 imputations and a maximum of 10 iterations. These choices strike a balance between computational efficiency and replicability. The *mice* package uses five donors by default, shown to be sufficient by van Buuren (2018).

For the glasso regularization, we utilized the *glasso* package (Friedman et al., 2019, Version 1.11). Model selection was performed in line with the standard implementation in *qgraph* (Version 1.9.8), using 100 penalty parameters and a minimum penalty value set to 0.01 times the maximum. Atan regularization was performed using the *GGMncv* package (Williams, 2021, Version 2.1.1) with 50 penalty parameters and the same ratio of 0.01. The hyperparameter for EBIC calculation was consistently set to $\gamma = 0.5$.

Results

This section is divided into results derived from $glasso_{EBIC}$ and $atan_{BIC}$. Given the extensive range of conditions and performance metrics, we report only the most central findings here. Additional figures and tables are available in the ESM, providing further details on results that are referenced but not discussed in depth in the main text. The ESM also includes an overview of key comparisons between the performance of the two approaches. In brief, the introduction of missingness has a more pronounced negative impact on sensitivity for $glasso_{EBIC}$, reducing it more strongly than for $atan_{BIC}$. In contrast, specificity is more negatively affected for $atan_{BIC}$, whereas it shows a slight improvement for $glasso_{EBIC}$. These deviations from full-data behavior are most evident under conditions with small sample sizes and high levels of missingness.

$glasso_{EBIC}$

Convergence

Overall, a single replication failed to converge (see Table ESM1 for an overview). This replication was excluded from all subsequent analyses.

Edge set

Figure 1 illustrates the density values across all replications for each condition. Performance can be evaluated in relation to both the complete data scenarios and the population densities, represented by dashed black lines. As expected, $glasso_{EBIC}$ yields increasing estimated densities with larger n, but these do not converge to the true population density due to systematic overestimation. As p grows, reaching the population density becomes more challenging, which is indicated by estimated densities well below population values even in complete data situations for the largest network. While the average density of all replications gets closer to the population density when keeping n constant and reducing the network size, the variability across replications increases. This is evident in conditions with and without missing values. With increased missingness, the estimated density of the networks gets smaller. However, this decline diminishes notably with larger observation numbers. The performance of the two-step EM and stacked MI approaches are closely aligned. In contrast, direct EM shows minor deviations, yet these deviations lack a discernible pattern. At a sample size of 400 observations, the performances of missing data handling approaches align closely with complete observations only when the missingness is minimal—otherwise, the resulting networks tend to be mostly empty. By contrast, with 800 observations, the methods generally perform adequately across all conditions. However, the direct EM approach faces challenges, especially with the medium and big networks.

Table 1 presents mean sensitivity and specificity for the various missing data handling techniques, alongside results for completely observed data. In general, sensitivity tends to be lower than specificity, as expected for an approach that, as described earlier, was designed to find a sparse solution in low-dimensional settings. Specificity exhibits a ceiling effect in conditions with small n, while sensitivity is low. As expected, this pattern reverses with a high number of observations, whereas, as indicated by the results reported for density, many edges are included, resulting in high sensitivity and low specificity. The size of the network also contributes to the tradeoff between sensitivity and specificity, with a higher network size leading to higher specificity. Increased missingness leads to a decrease in sensitivity and an increase in specificity, accompanied by greater variability within a condition (specific standard deviation values are available in Table ESM2). The impact on sensitivity is much more pronounced than on specificity. Yet, the effect of missingness diminishes with a higher number of observations. Handling methods demonstrate their ability to yield results similar to the complete data approach in almost all conditions with $n = 1600$, but only for 10% in the smallest network with $n = 400$. For 800 observations, a middle ground is observed where 10% and 20% missingness are similar to the complete data results for the smallest networks and

Table 1. Means of sensitivity and specificity across conditions using $glasso_{EBIC}$. Results of completely observed data shown under two-step EM for comparison.

		$n = 400$						$n = 800$						$n = 1600$					
		Stacked MI		Two-step EM		Direct EM		Sstacked MI		Two-step EM		Direct EM		Stacked MI		Two-step EM		Direct EM	
p	Prop.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
	0	–	–	0.914*	0.917*	–	–	–	–	0.995*	0.894*	–	–	–	–	1.000*	0.888*	–	–
	0.1	0.799	0.948	0.809	0.946	0.811	0.922	0.976	0.900	0.979	0.901	0.973	0.880	1.000	0.891	0.999	0.897	0.999	0.858
	0.2	0.602	0.966	0.616	0.966	0.554	0.949	0.941	0.911	0.946	0.909	0.912	0.873	0.992	0.885	0.994	0.887	0.981	0.830
8	0.3	0.213	0.992	0.241	0.990	0.225	0.987	0.815	0.918	0.813	0.914	0.670	0.911	0.975	0.867	0.976	0.871	0.912	0.819
	0	–	–	0.703*	0.970*	–	–	–	–	0.863*	0.940*	–	–	–	–	0.926*	0.927*	–	–
	0.1	0.594	0.981	0.606	0.980	0.539	0.985	0.833	0.941	0.831	0.943	0.817	0.945	0.909	0.926	0.909	0.927	0.908	0.927
	0.2	0.400	0.990	0.431	0.988	0.355	0.991	0.782	0.942	0.781	0.944	0.697	0.956	0.883	0.923	0.883	0.924	0.869	0.923
24	0.3	0.077	0.998	0.065	0.997	0.229	0.990	0.681	0.944	0.690	0.940	0.460	0.967	0.837	0.909	0.836	0.912	0.729	0.930
	0	–	–	0.415*	0.992*	–	–	–	–	0.639*	0.963*	–	–	–	–	0.729*	0.942*	–	–
	0.1	0.327	0.996	0.338	0.995	0.327	0.995	0.600	0.966	0.601	0.965	0.607	0.967	0.713	0.941	0.711	0.942	0.720	0.943
	0.2	0.233	0.997	0.256	0.995	0.244	0.995	0.544	0.969	0.547	0.967	0.501	0.978	0.686	0.939	0.683	0.941	0.701	0.942
48	0.3	0.119	0.999	0.131	0.998	0.179	0.994	0.450	0.973	0.451	0.965	0.361	0.979	0.635	0.933	0.632	0.934	0.581	0.959

Abbreviations: Prop.: Proportion of missing data. Sens.: Sensitivity. Spec.: Specificity.

*Results without missing values are computed by the complete data approach but presented in the column concerning two-step EM.

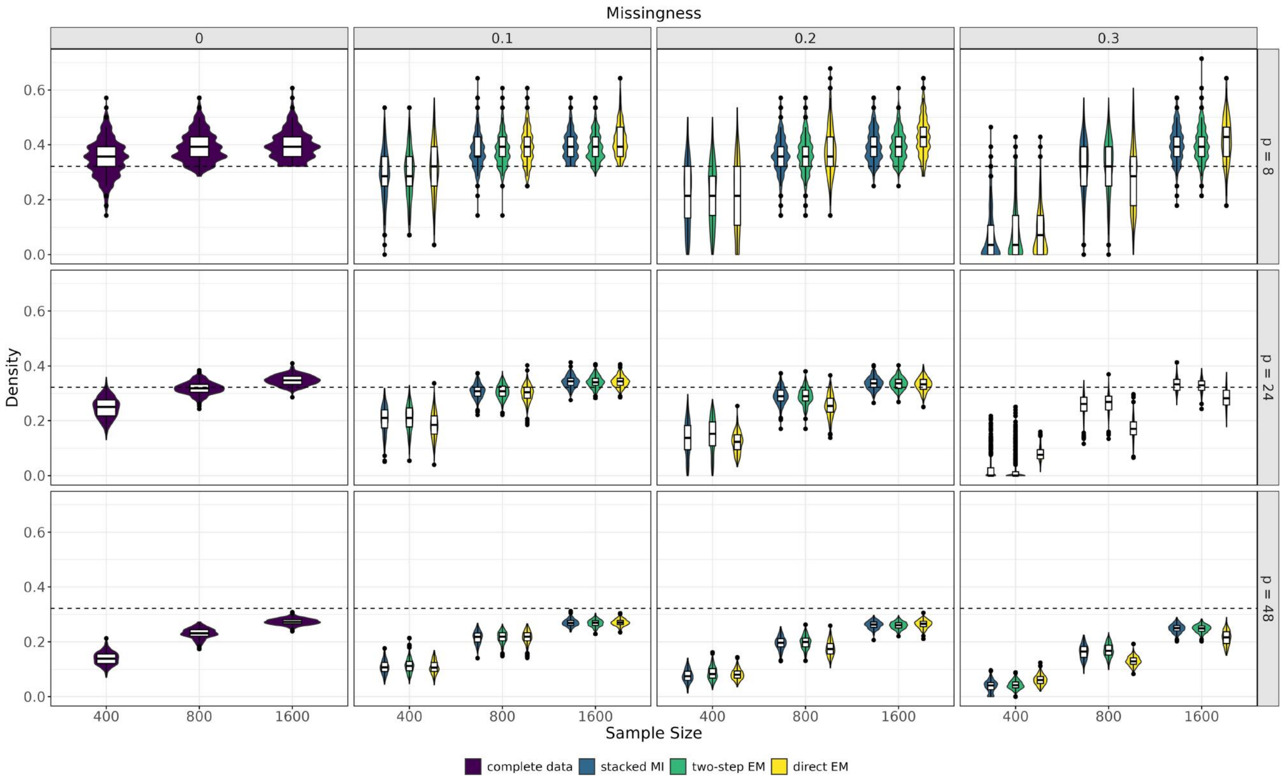

Figure 1. Density values of all replications for all conditions using $glasso_{EBIC}$. Columns vary different degrees of missingness, while rows vary network size. Dashed lines represent the true densities in the population networks.

Table 2. Means of sensitivity and specificity across conditions using $atan_{BIC}$. Results of completely observed data shown under two-step EM for comparison. Values for two-step EM are not available for the condition with 48 nodes, 400 observations, and 30% missing rate due to non-convergence.

		$n = 400$						$n = 800$						$n = 1600$					
		Stacked MI		Two-step EM		Direct EM		Stacked MI		Two-step EM		Direct EM		Stacked MI		Two-step EM		Direct EM	
p	Prop.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
	0	–	–	0.857*	0.958*	–	–	–	–	0.970*	0.980*	–	–	–	–	0.997*	0.990*	–	–
	0.1	0.816	0.955	0.812	0.954	0.797	0.967	0.944	0.970	0.941	0.970	0.931	0.977	0.992	0.986	0.993	0.986	0.988	0.990
	0.2	0.782	0.926	0.789	0.917	0.729	0.962	0.902	0.956	0.910	0.952	0.867	0.975	0.976	0.977	0.980	0.976	0.954	0.988
8	0.3	0.726	0.864	0.738	0.815	0.597	0.944	0.845	0.922	0.849	0.901	0.778	0.961	0.942	0.940	0.946	0.934	0.876	0.978
	0	–	–	0.546*	0.953*	–	–	–	–	0.642*	0.968*	–	–	–	–	0.729*	0.977*	–	–
	0.1	0.518	0.939	0.518	0.939	0.474	0.955	0.614	0.957	0.612	0.958	0.572	0.970	0.701	0.972	0.700	0.973	0.662	0.980
	0.2	0.491	0.908	0.499	0.890	0.422	0.946	0.582	0.934	0.578	0.935	0.518	0.962	0.669	0.955	0.668	0.955	0.606	0.976
24	0.3	0.521	0.773	0.629	0.561	0.360	0.926	0.538	0.875	0.537	0.864	0.446	0.935	0.625	0.911	0.623	0.908	0.552	0.940
	0	–	–	0.346*	0.954*	–	–	–	–	0.414*	0.965*	–	–	–	–	0.488*	0.973*	–	–
	0.1	0.327	0.941	0.325	0.937	0.306	0.958	0.398	0.953	0.396	0.953	0.372	0.968	0.468	0.965	0.465	0.966	0.441	0.977
	0.2	0.353	0.852	0.406	0.761	0.284	0.949	0.372	0.928	0.369	0.923	0.346	0.960	0.443	0.945	0.441	0.945	0.410	0.970
48	0.3	0.334	0.869	–	–	0.265	0.935	0.356	0.868	0.467	0.670	0.322	0.942	0.416	0.881	0.412	0.883	0.384	0.948

Abbreviations: Prop.: Proportion of missing data. Sens.: Sensitivity. Spec.: Specificity.
*Results without missing values are computed by the complete data approach but presented in the column concerning two-step EM.

10% for the medium and big networks. Examining the distinctions between missing handling methods, two-step EM exhibits higher sensitivity with 400 observations, while specificity remains comparable to the others. With 800 and 1600 observations, minimal differences exist between two-step EM and MI. Direct EM occasionally shows marginally better sensitivity, but it shows lower specificity and higher variety in results, making it less preferable.

Parameter estimation

Figure 2 illustrates the mean raw biases in correctly estimated non-zero partial correlations for conditions with 400 and 800 observations. Consistent with the expectations for $glasso_{EBIC}$, negative biases are observed across all conditions with 400 observations (depicted in Subfigure A). As seen previously for the estimated densities, the average bias gets closer to zero as n is held constant and the network size is reduced, but the variability across replications increases. Bias for completely observed data remains only very slightly below zero with minor deviations between replications. However, the introduction of missing values amplifies this bias, even with only 20% data missing. Furthermore, the variability in biases across replications raises concerns about the consistency and trustworthiness of the results. The most pronounced biases in single replications arise from instances where only a few partial correlations are correctly estimated as non-zero, and these are either small or incorrectly signed. Comparing the handling methods, all missing handling approaches demonstrate similar performance with 10% and 20 % missingness. This trend persists with 30% missingness for stacked MI and two-step EM. However, direct EM shows superior performance in these cases for the medium and big network sizes.

For conditions with 800 observations (depicted in Subfigure B), although partial correlations retain negative biases, these are, as expected, considerably closer to zero. The missing handling methods can manage up to 20 % missingness, yielding results akin to completely observed data. However, with 30% missingness, while the average performance remains comparable to the completely observed data for the small and medium networks, certain replications display significant biases, rendering the results of single replications questionable. Both stacked MI and two-step EM exhibit congruent outcomes, while direct EM presents marginally varied results, albeit without a discernible pattern. In the largest sample condition ($n = 1600$, Figure ESM1), the trends observed with increasing sample size persist. Negative biases further decrease, and the discrepancies between replications diminish, rendering even the 30% missingness scenario feasible. Additionally, the disparities between MI, two-step EM, and direct EM become minimal, indicating that the handling methods perform comparably at this sample size.

Network statistics

Figure 3 depicts the correlations between strength values from estimated and population networks for conditions with 400 and 800 observations. Notably, this metric shows the largest discrepancies between completely observed cases and those with missing data. As network size increases, the correlations tend to decrease, indicating less reliable identification of strongly connected nodes. With 400 observations (depicted in Subfigure A), the capacity to roughly discern the nodes with highest strength values remains

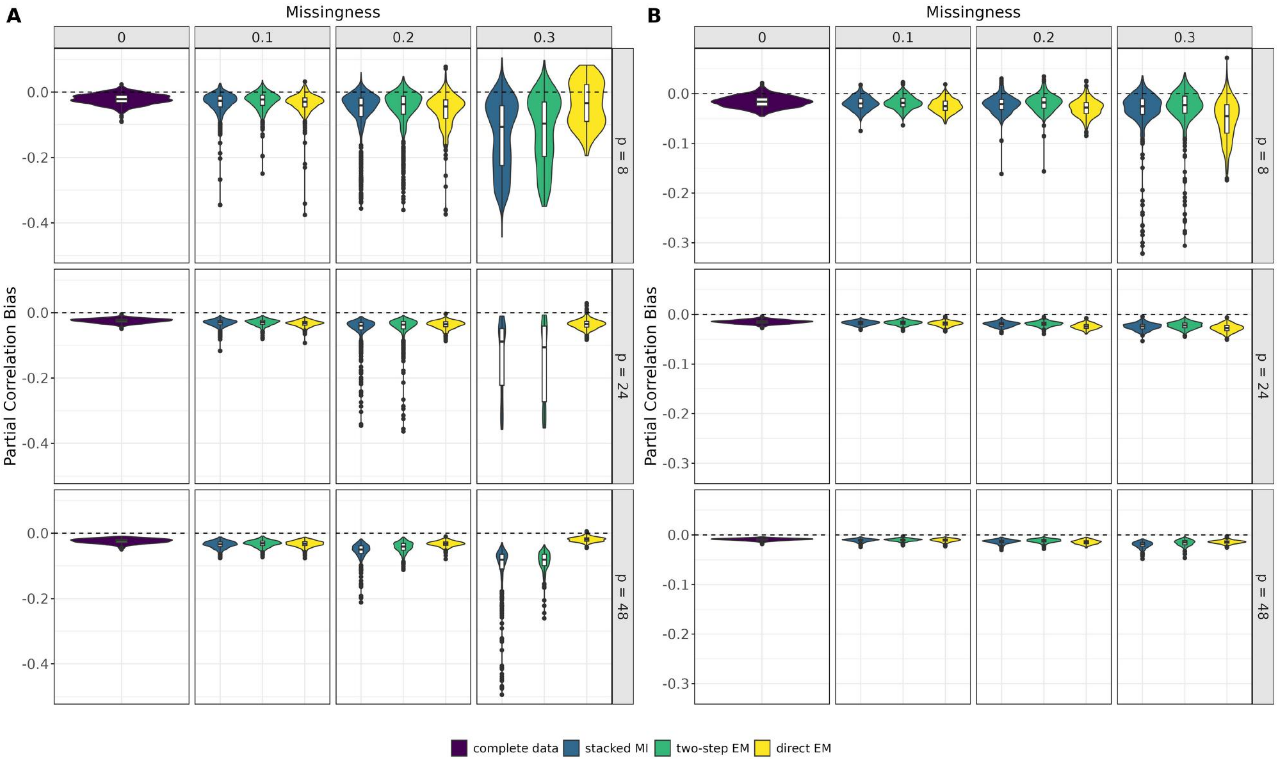

Figure 2. Mean raw bias of correctly identified non-zero partial correlations per replication using $glasso_{EBIC}$. Conditions with **A** 400 and **B** 800 observations. Columns vary different degrees of missingness, while rows vary network size. Scaling of y-axis differs between **A** and **B**.

intact for the largest network configurations, provided no data is missing. Outcomes remain satisfactory with 10% missingness but, as we anticipated, an increase in missingness leads to a decline in correlation. Starting with 20% missingness, the variance across individual replications becomes too pronounced for reliable conclusions. Both stacked MI and two-step EM yield consistently analogous results, though stacked MI returns smaller correlations for single replications. In contrast, direct EM consistently underperforms across most scenarios, including situations with significant missing data and expansive networks.

In line with our expectations, when the sample size is increased to 800 observations (depicted in Subfigure B), both stacked MI and two-step EM closely approximate results from completely observed data in the presence of 10% and 20% missingness. At a 30% missing rate, their performance slightly weakens but remains within an acceptable spectrum. The variability in results decreases, with instances of small correlations becoming rare. However, an exception arises in the context of 30% missingness within the smallest network, where correlations occasionally approach zero. Direct EM underperforms in direct comparison even with a 20% missing rate for 800 observations. With 1600 observations (Figure ESM2), overall performance continues to improve. It is noteworthy that no individual replication records a correlation below 0.6, even under 30% missingness for stacked MI and two-step EM. However, direct EM still exhibits suboptimal outcomes, characterized by substantial variations across replications, especially at the 30% missingness rate.

$atan_{BIC}$

Convergence

For completely observed data sets, there were no issues with convergence. However, when attempting to employ $atan_{BIC}$ in conjunction with missing data handling techniques, complications arose in conditions with $n = 400$, as well as conditions with $n = 800$ and $p = 48$. Non-convergence with two-step EM occurred during the second step, indicating that the missing data handling in the first step, which was conducted using the *lavaan* package was completed successfully. However, this initial step occasionally yielded covariance matrices that indicated nearly perfect correlations between all variables. This resulted in non-positive definite matrices after applying atan regularization, which ultimately led to the termination of computations in the second step. Non-convergence

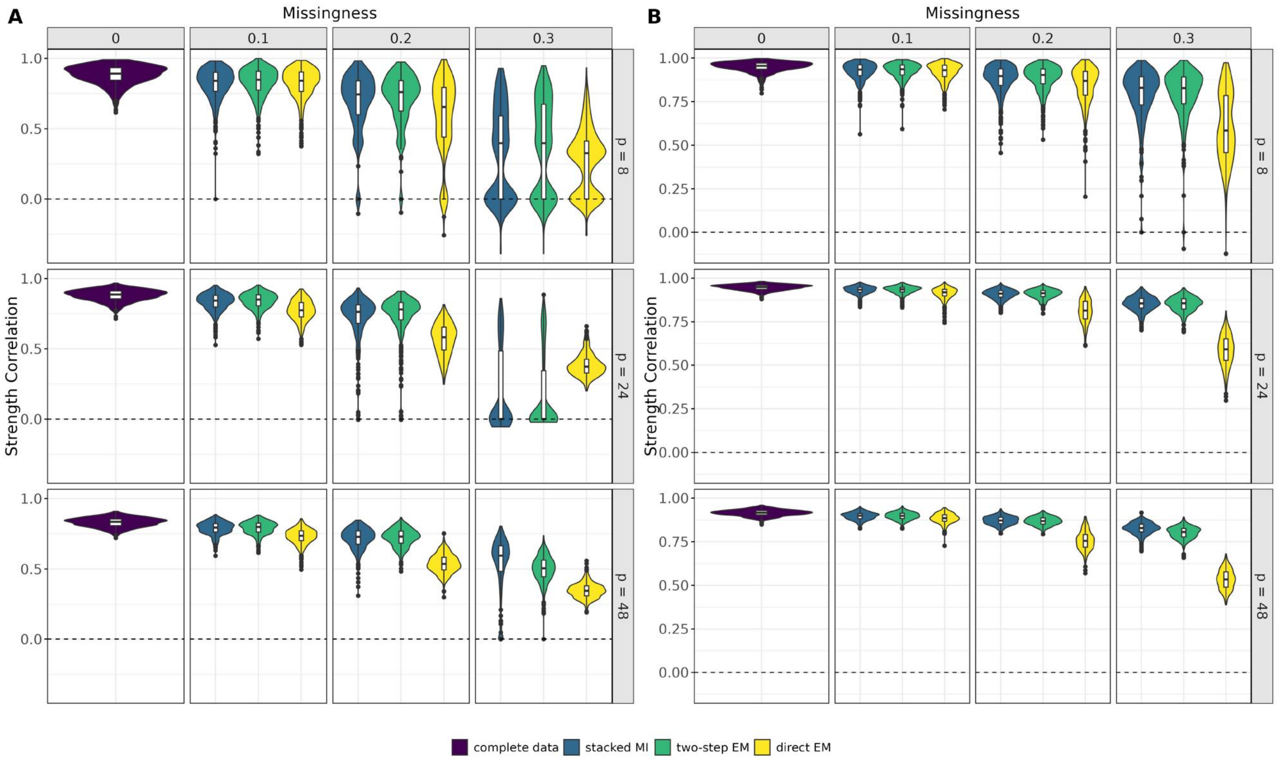

Figure 3. Correlations of strength values between estimated and population networks using $glasso_{EBIC}$. Conditions with **A** 400 and **B** 800 observations. Columns vary different degrees of missingness, while rows vary network size. Scaling of y-axis differs between **A** and **B**.

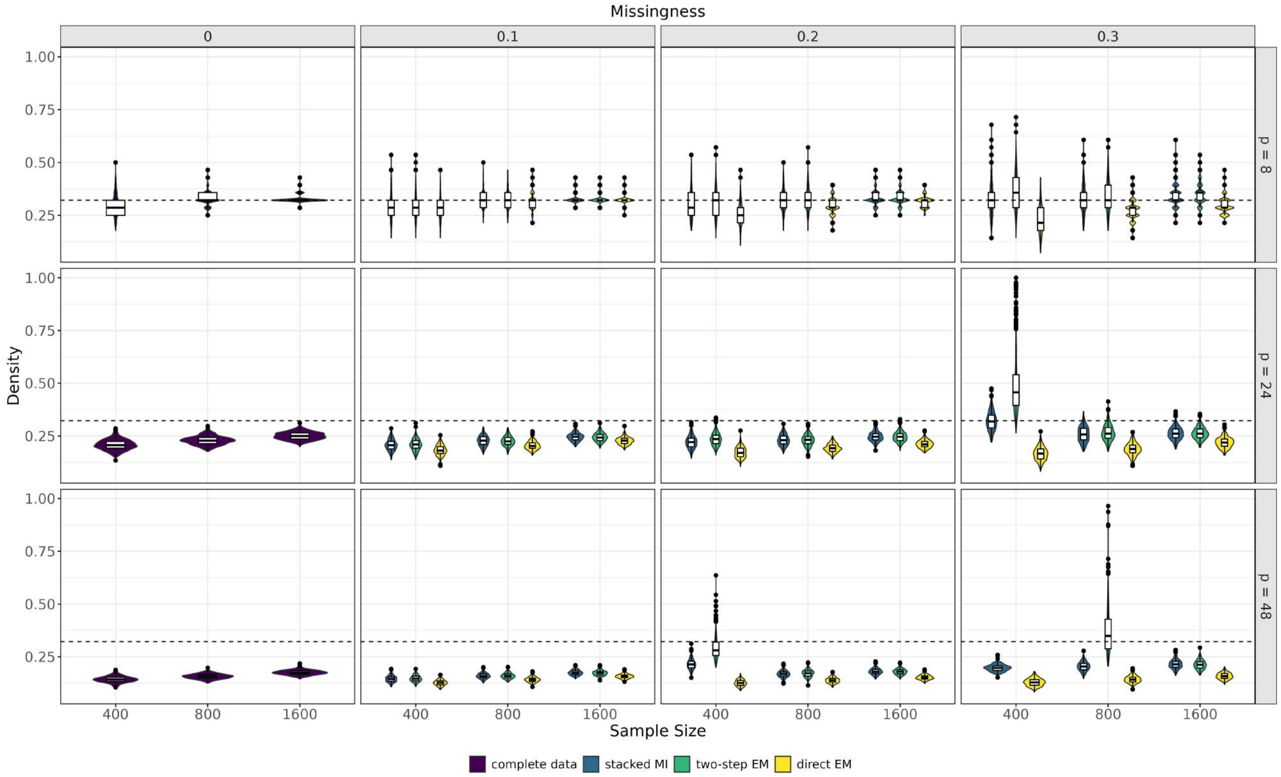

Figure 4. Density values of all replications for all conditions using $atan_{BIC}$. Columns vary different degrees of missingness, while rows vary network size. Dashed lines represent the true densities in the population networks. Values for two-step EM are not available for the condition with 48 nodes, 400 observations, and 30% missing rate due to non-convergence.

was particularly evident for 30% missingness, reaching a point where all replications failed with $p = 48$ and $n = 400$. Even for $n = 800$, almost half of the replications failed for the big network. Direct EM exhibited fewer and less severe convergence issues, primarily affecting a small subset of cases—specifically, the largest network with the smallest sample size, where about one-fifth of the replications failed. It is important to note that a replication was classified as a failure if convergence was not achieved for all penalty parameters tested. In contrast, stacked MI demonstrated successful convergence across all conditions. Detailed convergence rates are provided in the ESM (Table ESM3). Replications that failed to converge were excluded from all subsequent analyses.

Edge set

Figure 4 depicts the density values for all replications of every condition. In line with our expectations, the estimated density approaches the population value in the completely observed conditions with $n = 800$ and $n = 1600$ in the smallest network. When considering conditions with missing values, larger sample sizes tend to result in networks that are more closely aligned with their counterparts from completely observed conditions. Specifically, for 400 observations, the networks are comparable to the completely observed ones only when missingness is at 10%. With 800 observations, this similarity extends to 20% missingness. However, it is worth noting that even conditions with completely observed data often yield a very sparse structure for the big network. For the two-step EM approach, reliable results with 30% missingness are only observed when using 1600 observations. In contrast, both direct EM and stacked MI seem stable with 30% missingness in larger networks with 400 and 800 observations. However, this stability is somewhat misleading as these methods often produce nearly empty network structures. Overall, stacked MI offers the most stable outcomes, while the two-step EM approach occasionally comes closer to the density estimated by the complete data sets, especially in conditions with minimal missingness. Direct EM typically lags slightly behind both two-step EM and stacked MI in terms of average performance, while producing relatively stable results across its replications.

Table 2 presents a comparative overview of mean sensitivity and specificity across conditions with and without missing data. For cases with complete data, specificity is generally high, while sensitivity increases with n. Notably, there is no clear tradeoff between these two metrics. However, even with 1600 observations,

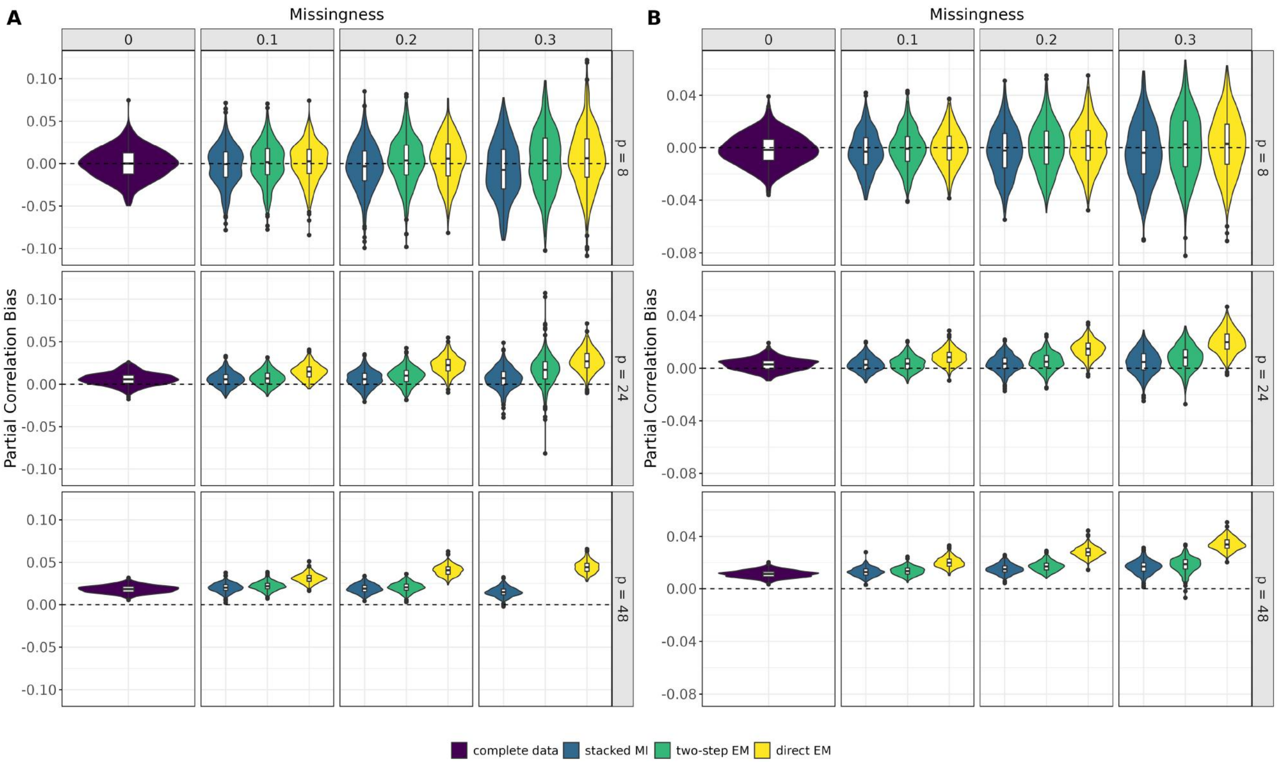

Figure 5. Mean raw bias of correctly identified non-zero partial correlations per replication using $atan_{BIC}$. Conditions with **A** 400 and **B** 800 observations. Columns vary different degrees of missingness, while rows vary network size. Scaling of y-axis differs between **A** and **B**. Values for two-step EM are not available for the condition with 48 nodes, 400 observations, and 30% missing rate due to non-convergence.

complete data fails to achieve high sensitivity values for larger networks, indicating a lack of power. The introduction of missingness results in a decline in both sensitivity and specificity, with a more pronounced effect on sensitivity. This phenomenon is accompanied by an increase in variability across replications within the same condition (specific standard deviation values are provided in Table ESM4). The observed effects are less prominent in smaller networks and conditions with a higher number of observations. In line with our expectations, an increase in sample size leads to an increase in sensitivity while maintaining high specificity in conditions with and without missingness. For 400 observations, missing data handling methods demonstrate the ability to perform comparably to the complete data approach in conditions with 10% missingness and 400 observations. In the case of 1600 observations, results with 20% missingness are similar to complete data. A more substantial increase in observations would be required to mitigate the impact of 30% missingness.

The comparison between missing handling methods reveals nuanced patterns. For data sets with 400 observations, two-step EM demonstrates significantly greater variability across replications compared to direct EM and stacked MI, especially with a 30% missingness rate. Direct EM outperforms the other two approaches with regards to specificity in these conditions, although values obtained with stacked MI are close and the latter shows higher sensitivity. With 400 observations, two-step EM is competitive in both sensitivity and specificity only at a 10% missingness rate. With 800 observations, two-step EM continues to face challenges with variability across replications, especially as missingness increases and the network size grows. In contrast, stacked MI and direct EM exhibit performance closer to complete data, with stacked MI being more sensitive and direct EM more specific. As the sample size increases to 1600 observations, distinctions between the approaches diminish. Yet, a consistent trend emerges wherein both stacked MI and two-step EM prioritize sensitivity in contrast to direct EM's inclination toward specificity.

Parameter estimation

The mean raw bias of correctly identified non-zero partial correlations per replication is illustrated in Figure 5 for data sets with 400 and 800 observations. On average, the bias approaches zero, but is skewed positive, indicating an overestimation of partial correlations. This positive bias intensifies with increasing missingness and larger networks. For data sets with 400 observations (depicted in Subfigure A), there is a

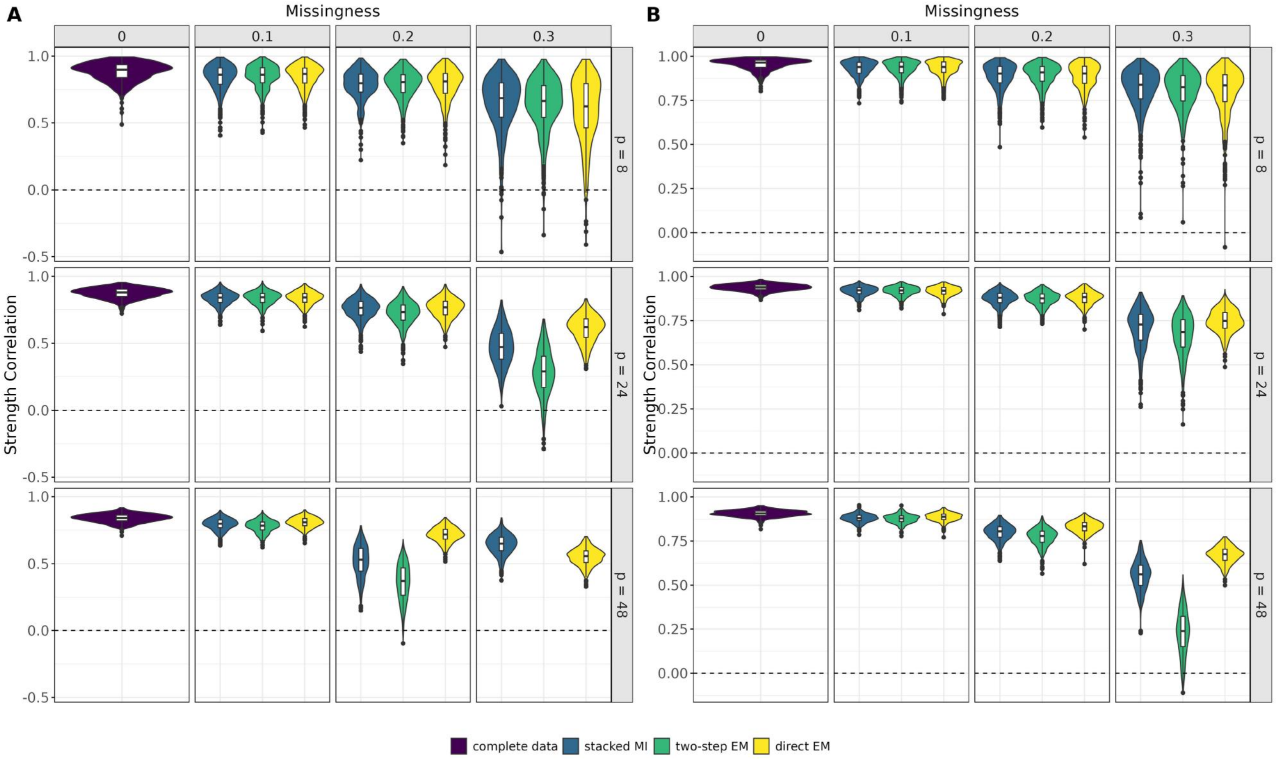

Figure 6. Correlations of strength values between estimated and population networks using $atan_{BIC}$. Conditions with **A** 400 and **B** 800 observations. Columns vary different degrees of missingness, while rows vary network size. Scaling of y-axis differs between **A** and **B**. Values for two-step EM are not available for the condition with 48 nodes, 400 observations, and 30% missing rate due to non-convergence.

systematic positive bias evident in completely observed scenarios for both $p = 24$ and $p = 48$. While the average bias decreases with a smaller network and constant sample size, the variability between replications increases (as already seen in the $glasso_{EBIC}$ results). Stacked MI consistently mirrors the performance of completely observed data across all conditions. Two-step EM encounters challenges in individual replications, particularly evident with a 30% missingness rate. Direct EM registers the highest positive bias, with its estimates closely aligning to the other two handling techniques only under minimal missingness or within smaller networks.

With 800 observations (depicted in Subfigure B), the systematic bias diminishes, as we anticipated, but persists, especially for the larger networks. Stacked MI aligns closely with completely observed data on average, yet it exhibits a broader range across replications, particularly at the 30% missingness level. While two-step EM's average performance mirrors that of MI, it demonstrates an even greater variability. Direct EM continues to exhibit the most pronounced positive bias but shows a similar variability across replications. Using a sample size of $n = 1600$ (Figure ESM3) results in a substantial reduction in bias, with virtually no systematic bias observed at $p = 48$. In addition, variability between replications decreases, although it remains slightly higher in conditions with missing data compared to complete data. The relative performance patterns among the missing data handling techniques persist, with both stacked MI and two-step EM offering similar results, while direct EM consistently exhibits the highest bias.

Network statistics

Correlations derived from strength values of estimated and population networks with 400 and 800 observations are illustrated in Figure 6. Again, while the average correlation is closer to 1 with a constant sample size and a smaller network, the variation across replications also increases. Direct EM consistently exhibited the highest correlations with 400 observations (depicted in Subfigure A), performing similarly to the complete data conditions even when subjected to 10% and 20% missingness. While there were isolated instances of poor correlation, such anomalies were also observed using the complete data set. As expected, the discrepancies in the correlations became more pronounced with increased missingness, raising concerns about the reliability of the results when 20% of data were missing in the big network. Stacked MI showcased performance only marginally inferior to the complete data set and direct EM in the 10% missingness scenario. As seen in other evaluation criteria, two-step EM exhibited limitations as the level of missingness escalated.

In line with our expectations, 800 observations (Subfigure B) generally led to higher correlations. A notable ceiling effect was observed with the completely observed data in the small network, correlations for the medium and big networks without missings never dropped below $r = .75$. Techniques for handling missing data demonstrated enhanced resilience, effectively managing both 10% and 20% missingness. Nevertheless, certain replications exhibited outlier behavior. In conditions with 10% and 20% missingness, stacked MI and the two-step method exhibited performance on par with direct EM. A modest advantage for direct EM was discerned with $p = 48$ at 20% missingness. However, the deviations for the 30% missingness scenario, though diminished compared to situations with 400 observations, were still substantial, calling into question the reliability of the returned strength values. Conditions with 1600 observations (Figure ESM4) showed even better results, preserving the previously observed patterns. The magnitude of deviations also decreased for the 30% missingness condition. Nevertheless, sporadic replications still recorded correlations hovering around $r = .5$. Direct EM exhibited improved correlations compared to stacked MI and two-step EM in scenarios with $p = 48$ and a 30% missingness rate.

Discussion

The aim of the present study was to investigate the performance of a stacked MI approach in cross-sectional, psychological network analysis, contrasting it with two distinct EM approaches. This was done for both $glasso_{EBIC}$ and $atan_{BIC}$. As the computation of model selection criteria was standardized across the missing data handling methods, the recommendations for preferred techniques, detailed later, differ from earlier findings which employed different calculations (Falk & Starr, 2023) and which relied on the default implementations of R packages (Nehler & Schultze, 2024b). It also necessitates a reevaluation of required sample sizes in cases with missing values described by Nehler and Schultze (2024b). In the study presented here, higher levels of missingness generally led to less dense networks with decreases in sensitivity, which may seem less desirable. However, these results can be viewed as more representative of the actual amount of available information.

In line with the standard implementation of $glasso_{EBIC}$ without an additional control for the false positive rate, a negative bias was observed, and as the sample size increased, the network became denser with a noticeable decline in specificity. This is consistent with the findings of Williams et al. (2019), who showed that consistency in model selection is lacking when the network structure is not extremely sparse, which is the case in our simulation. This effect may change by incorporating approaches designed for controlling the false positive rate (e.g., Lafit et al., 2019). Conducting network analysis *via* $glasso_{EBIC}$ with a data set of 400 individuals and a missingness exceeding 10% is not recommended. Generally, conditions with 800 observations remained robust against missing data up to 30%. However, there were outliers in strength correlations, particularly in the small network. When scaling up to 1600 observations, the decline in specificity persisted even in the presence of missingness. Thus, this implementation of $glasso_{EBIC}$ does not converge to the true model with perfect sensitivity and specificity but rather continues to add more edges with increased sample size, making the approach less favorable in these situations regardless of the missing data handling technique employed. In terms of comparing the handling techniques, direct EM exhibited marginally inferior performance in certain scenarios and significantly poorer outcomes in others, suggesting it may not be the optimal choice across the board. This may be due to the intentional model misfit introduced by regularization in each iteration of the direct EM approach, which could increase bias—as demonstrated by Lee and Shi (2021) in the context of SEM. Both stacked MI and two-step EM exhibited similar performance across all evaluated conditions. As likelihood based approaches are acknowledged as the gold standard in previous literature (van Buuren, 2018), with $glasso_{EBIC}$, the two-step EM appears to be a reasonable recommendation. Additionally, it usually offers computational advantages in terms of speed. Nevertheless, stacked MI presents a viable alternative, especially when other analyses are planned for the same data set.

For $atan_{BIC}$, the relationship between density and n is less pronounced, as described by Williams (2020). While 400 observations are sufficient for smaller networks, larger networks necessitate a larger sample size. In particular, even with 1600 observations, the performance remains suboptimal for a 48-node network. However, this network size could be considered an edge case based on a review of applied studies (Wysocki & Rhemtulla, 2021). The required sample sizes for handling missing data vary: 400 observations suffice for a 10% missingness rate, 800 for 20%, and 1600 for 30%. Nonetheless, some replications exhibited outlier behavior, suggesting potential pitfalls for individual researchers in achieving accurate results. When contrasting missing data handling techniques, more nuanced recommendations are necessary for $atan_{BIC}$ than for $glasso_{EBIC}$. Specifically, for 400 observations, direct EM may be recommended for identifying nodes with the highest strength. However, caution is advised due to the heightened risks of false negatives and increased bias in estimated nodes. The two-step EM exhibits robust performance with larger data sets but demonstrates inconsistency with fewer observations, higher node counts, or elevated missingness rates. Stacked MI consistently provides the most stable performance across various conditions, making it the recommended approach when using $atan_{BIC}$.

Notably, the discrepancies in performance of the missing handling techniques between convex and nonconvex regularization in our simulations suggest that a universally optimal technique remains unclear. Consequently, the choice of a method should depend on the specific context of each applied study.

While our results indicate the necessity of substantial sample sizes for both $glasso_{EBIC}$ and $atan_{BIC}$, a review of applied studies in cross-sectional network analysis revealed that almost half of the studies encompassed more than 500 observations (Wysocki & Rhemtulla, 2021). It is crucial to acknowledge that in cases with a smaller sample size, simply opting for a smaller network is not always the solution for achieving a reliable outcome. Although accuracy tends to increase on average when comparing results between conditions with a larger and smaller network while maintaining the same sample size, the discrepancies between replications within the same conditions also tend to escalate. This phenomenon is not exclusive to conditions with missing values—it also occurs in those with completely observed data sets.

Limitations and future directions

The contrasting outcomes regarding the best missing handling technique from prior evaluations (Falk & Starr, 2023; Nehler & Schultze, 2024b) underscored the critical influence of decisions in model selection with missing data. In particular, the determination of sample size and the method employed for computing the log-likelihood were pivotal in computing both EBIC and BIC. Generally, using individual log-likelihoods and calculating average sample size resulted in the selection of less dense networks compared to the previous study of Nehler and Schultze (2024a). Nevertheless, conducting a dedicated study comparing these approaches could provide a more nuanced understanding of their respective impacts. Notably, for EBIC, the implications extend further, including the challenge of selecting an appropriate value for the hyperparameter γ. Given these complexities, the value of selecting the best model based on methods like cross-validation (Krämer et al., 2009), which was originally the main suggestion of Städler and Bühlmann (2012), warrants thorough evaluation in a comparative study. While Falk and Starr (2023) include cross-validation, it is solely implemented for the direct EM approach, where it demonstrates encouraging outcomes. Moreover, if the primary aim is to compare network estimation methods, future research should also consider varying the information criteria, applying the EBIC in combination with atan and the BIC with glasso, to implement a fully crossed experimental design.

The present study used psychological data to construct population networks with the objective of emulating realistic scenarios. Although the observed effects remained consistent when comparing transitions from small to medium and medium to large network size, these findings warrant validation using other realistic psychological structures. It is worth noting that our study maintained the same density for all population networks, representing a case of sparse to medium density. Future investigations might benefit from varying the density, as advocated by studies such as Williams (2020). The varied conditions within the simulation were selected with the understanding that the results from different missing data handling methods would likely converge, particularly for the two-step EM and stacked MI approaches, which are asymptotically equivalent when assumptions hold. The primary aim was to identify subtle discrepancies and validate the conceptual framework of these methods. Future simulation studies could enhance differentiation by exploring for example deviations from distributional assumptions.

An alternative approach to integrating MI in network analysis is grouping, which, unlike stacked MI, does not combine the data sets but instead ensures that each parameter is either consistently included or excluded across all imputations. This makes grouped MI conceptually less similar to the two-step EM procedure than stacked MI, although it may be less easily generalized due to the inherent specificity required for different types of model estimation and selection. In the context of regularized regression, the method has shown promising results (Chen & Wang, 2013). However, in network analysis without regularization, its initial implementation demonstrated weaker performance (Nehler & Schultze, 2024a). This raises the question of whether grouped MI might yield better results in regularized network models. Another approach, discussed in the statistical and machine learning literature, involves modifying the regularization procedure to directly accommodate missing data (Loh & Wainwright, 2012, 2015). This method adapts the estimator itself, in contrast to the approaches presented in our manuscript, where regularization is either integrated into missing data handling (direct EM) or addressed separately (two-step EM and stacked MI). While the cited studies have demonstrated the effectiveness of these modified estimators, future research should evaluate their performance within the context of psychological networks and compare them to the methods presented in this manuscript.

On a more general note, investigating the phenomenon of increased variability across replications while keeping a consistent number of observations but smaller network sizes presents an important avenue for future research. Reducing the number of nodes could be seen as an intuitive approach for applied researchers to enhance the reliability of their results in cases with a low number of available observations. However, our results suggest that this approach may not be as straightforward as it seems. Thus, the underlying reasons for this variability warrant thorough investigation. Additionally, it is essential to explore whether this effect is specific to regularization techniques or also occurs in non-regularized methods (e.g., Williams et al., 2019). Examining a broader range of network estimation and selection techniques will also further strengthen the understanding of how well the findings on missing data handling generalize, building on the initial step taken in this study.

Conclusion

The present study demonstrated that stacked MI and EM algorithms for cross-sectional network analysis under standardized model estimation and selection exhibited similar, though not equivalent, performance. The analyses indicated that when using $glasso_{EBIC}$, the two-step EM approach is recommended for handling missing values. In contrast, the choice of missing data handling technique for network analysis with $atan_{BIC}$ is more nuanced, with stacked MI generally being the most stable.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by any funding.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Augugliaro, L., Sottile, G., Wit, E. C., & Vinciotti, V. (2023). Cglasso: An R package for conditional graphical lasso inference with censored and missing values. *Journal* of Statistical Software, 105(1), 58. https://doi.org/10.18637/jss.v105.i01

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747–3752. https://doi.org/10.1073/pnas.0400087101

Bates, D., Maechler, M., Jagan, M. (2023). *Matrix: Sparse and dense matrix classes and method*. https://CRAN.R-project.org/package=Matrix

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T. W., & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892–903. https://doi.org/10.1037/abn0000446

Brown, L. L., Acevedo, B., & Fisher, H. E. (2013). Neural correlates of four broad temperament dimensions: Testing predictions for a novel construct of personality. *PloS One*, *8*(11), e78734. https://doi.org/10.1371/journal.pone.0078734

Burger, J., Isvoranu, A.-M., Lunansky, G., Haslbeck, J. M. B., Epskamp, S., Hoekstra, R. H. A., Fried, E. I., Borsboom, D., & Blanken, T. F. (2023). Reporting standards for psychological network analyses in cross-sectional data. *Psychological Methods*, *28*(4), 806–824. https://doi.org/10.1037/met0000471

Chen, Q., & Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, *32*(21), 3646–3659. https://doi.org/10.1002/sim.5783

Core Team, R. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (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. https://doi.org/10.1016/j.jrp.2014.07.003

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713–740. https://doi.org/10.1177/0013164401615001

Epskamp, S. (2016). *Regularized Gaussian psychological networks: Brief report on the performance of extended BIC model selection*. arXiv. https://doi.org/10.48550/arXiv.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. https://doi.org/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), 1–18. https://doi.org/10.18637/jss.v048.i04

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Falk, C. F., & Starr, J. (2023). *Regularized cross-sectional network modeling with missing data: A comparison of methods*. PsyArXiv. https://doi.org/10.31219/osf.io/dk6zv

Fan, J., Xue, L., & Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, *42*(3), 819–849. https://doi.org/10.1214/13-aos1198

Fisher, H. E., Rich, J., Island, H. D., & Marchalik, D. (2010). The second to fourth digit ratio: A measure of two hormonally-based temperament dimensions. *Personality and Individual Differences*, *49*(7), 773–777. https://doi.org/10.1016/j.paid.2010.06.027

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, *23*, 604–612. https://arxiv.org/abs/1011.6640

Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O. J., Epskamp, S., Tuerlinckx, F., Carr, D., & Stroebe, M. (2015). From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology*, *124*(2), 256–265. https://doi.org/10.1037/abn0000028

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

Friedman, J., Hastie, T., & Tibshirani, R. (2019). Glasso: Graphical lasso-estimation of Gaussian graphical models. *R Package Version* 1.11. https://CRAN.R-project.org/package=glasso

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Isvoranu, A.-M., & Epskamp, S. (2023). Which estimation method to choose in network psychometrics? Deriving guidelines for applied researchers. *Psychological Methods*, *28*(4), 925–946. https://doi.org/10.1037/met0000439

Jefferies, P., Höltge, J., Fritz, J., & Ungar, M. (2023). A cross-country network analysis of resilience systems in young adults. *Emerging Adulthood (Print)*, *11*(2), 415–430. https://doi.org/10.1177/21676968221090039

Jia, F., & Wu, W. (2023). A comparison of multiple imputation strategies to deal with missing nonnormal data in structural equation modeling. *Behavior Research Methods*, *55*(6), 3100–3119. https://doi.org/10.3758/s13428-022-01936-y

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, *42*(4), 371–404. https://doi.org/10.3102/1076998616687084

Krämer, N., Schäfer, J., & Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics*, *10*(1), 384. https://doi.org/10.1186/1471-2105-10-384

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lafit, G., Tuerlinckx, F., Myin-Germeys, I., & Ceulemans, E. (2019). A partial correlation screening approach for controlling the false positive rate in sparse Gaussian graphical models. *Scientific Reports*, *9*(1), 17759. https://doi.org/10.1038/s41598-019-53795-x

Lang, K. M., & Little, T. D. (2014). The supermatrix technique: A simple framework for hypothesis testing with missing data. *International Journal of Behavioral Development*, *38*(5), 461–470. https://doi.org/10.1177/0165025413514326

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

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods*, *26*(4), 466–485. https://doi.org/10.1037/met0000381

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

Loh, P.-L., & Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics*, *40*(3), 1637–1664. https://doi.org/10.1214/12-AOS1018

Loh, P.-L., & Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, *16*(1), 559–616. https://doi.org/10.5555/2789272.2789291

Lu, J.-X., Zhai, Y.-J., Chen, J., Zhang, Q.-H., Chen, T.-Z., Lu, C.-L., Jiang, Z.-L., Guo, L., & Zheng, H. (2023). Network analysis of internet addiction and sleep disturbance symptoms. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *125*, 110737. https://doi.org/10.1016/j.pnpbp.2023.110737

Mansueto, A. C., Wiers, R. W., van Weert, J. C. M., Schouten, B. C., & Epskamp, S. (2023). Investigating the feasibility of idiographic network models. *Psychological Methods*, *28*(5), 1052–1068. https://doi.org/10.1037/met0000466

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Miers, A. C., Weeda, W. D., Blöte, A. W., Cramer, A. O. J., Borsboom, D., & Westenberg, P. M. (2020). A cross-sectional and longitudinal network analysis approach to understanding connections among social anxiety components in youth. *Journal of Abnormal Psychology*, *129*(1), 82–91. https://doi.org/10.1037/abn0000484

Nehler, K. J., & Schultze, M. (2024a). *Handling missing values when using neighborhood selection for network analysis*. PsyArXiv. https://doi.org/10.31234/osf.io/qpj35

Nehler, K. J., & Schultze, M. (2024b). Simulation-based performance evaluation of missing data handling in network analysis. *Multivariate Behavioral Research*, *59*(3), 461–481. https://doi.org/10.1080/00273171.2023.2283638

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. https://doi.org/10.1016/j.socnet.2010.03.006

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987a). Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 542–543. https://doi.org/10.1080/01621459.1987.10478461

Rubin, D. B. (1987b). *Multiple imputation for survey nonresponse*. Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.1080/01621459.1996.10476908

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 477–497. https://doi.org/10.1080/10705510903008238

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Städler, N., & Bühlmann, P. (2012). Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, *22*(1), 219–235. https://doi.org/10.1007/s11222-010-9219-7

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. https://doi.org/10.1177/0962280206074463

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681–694. https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6<681::AIDSIM71>3.0.CO;2-R

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. https://doi.org/10.1080/10629360600810434

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van Zyl, C. (2021). A network analysis of the General Health Questionnaire. *Journal of Health Psychology*, *26*(2), 249–259. https://doi.org/10.1177/1359105318810113

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Vink, G., & Van Buuren, S. (2013). Multiple imputation of squared terms. *Sociological Methods & Research*, *42*(4), 598–607. https://doi.org/10.1177/0049124113502943

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*(3), 699–718. https://doi.org/10.1177/0049124117747303

Wan, Y., Datta, S., Conklin, D., & Kong, M. (2015). Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation*, *85*(9), 1902–1916. https://doi.org/10.1080/00949655.2014.907801

Wang, Y., & Zhu, L. (2016). Variable selection and parameter estimation with the atan regularization method. *Journal of Probability and Statistics*, *2016*, e6495417–12. https://doi.org/10.1155/2016/6495417

Williams, D. R. (2020). *Beyond lasso: A survey of nonconvex regularization in Gaussian graphical models.* PsyArXiv. https://doi.org/10.31234/osf.io/ad57p

Williams, D. R., & Rast, P. (2020). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 187–212. https://doi.org/10.1111/bmsp.12173

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research*, *54*(5), 719–750. https://doi.org/10.1080/00273171.2019.1575716

Williams, D. R. (2021). GGMncv: Gaussian graphical models with nonconvex regularization. https://CRAN.R-project.org/package=GGMncv

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, *27*(17), 3227–3246. https://doi.org/10.1002/sim.3177

Wysocki, A. C., & Rhemtulla, M. (2021). On penalty parameter selection for estimating network models. *Multivariate Behavioral Research*, *56*(2), 288–302. https://doi.org/10.1080/00273171.2019.1672516

Zhang, X., & Savalei, V. (2023). New computations for RMSEA and CFI following FIML and TS estimation with missing data. *Psychological Methods*, *28*(2), 263–283. https://doi.org/10.1037/met0000445

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*(4), 1509–1533. https://doi.org/10.1214/009053607000000802

Appendix A

Additional information on the population networks

As described in the introduction, the consistency of the network analysis methods relies on different assumptions. To gain a deeper understanding of the data-generating networks, this appendix presents the distribution of partial correlations for all three sizes in Table A1. As noted in the methods section, the density is equal across network sizes, and therefore the number of zero partial correlations is the same. Overall, the number of larger partial correlations decreases with increasing network size, which is expected, as more variables are partialed out.

Table A1. Size of partial correlations in the population networks.

|Pcor|	p = 8	p = 24	p = 48
= 0	67.86%	67.75%	67.82%
> 0–0.1	0.00%	18.12%	25.71%
> 0.1–0.2	17.86%	10.87%	5.05%
> 0.2–0.3	7.14%	2.54%	0.71%
> 0.3–0.4	3.57%	0.00%	0.35%
> 0.4–1	3.57%	0.72%	0.35%

Abbreviations: \|Pcor\| = Absolute value of partial correlation.

Furthermore, Table A2 examines the detectability of the population networks to improve the interpretability of our results. Detectability is assessed by assuming an infinite sample size and providing the population covariance matrix directly to the estimation and selection procedures—either *glasso*$_{EBIC}$ or *atan*$_{BIC}$. The resulting network is then evaluated in terms of sensitivity and specificity relative to the true population network. As shown in the table, consistent recovery is possible for the network with 8 nodes. For the 24-node network, sensitivity remains nearly perfect for *glasso*$_{EBIC}$, but is slightly lower for *atan*$_{BIC}$. For the network with 48 nodes, both sensitivity and specificity decrease, but remain acceptable for *glasso*$_{EBIC}$. In contrast, for *atan*$_{BIC}$, sensitivity is relatively low, although specificity remains perfect.

Table A2. Investigation of detectability of the population networks.

	glasso$_{EBIC}$		*atan*$_{BIC}$	
p	Spec.	Sens.	Spec.	Sens.
8	1.000	1.000	1.000	1.000
24	1.000	0.989	1.000	0.944
48	0.983	0.898	1.000	0.647

Abbreviations: Sens.: Sensitivity. Spec.: Specificity.

Appendix B

Technical implementation of the direct EM

As described in the main text, the direct EM is implemented following the description provided by Städler and Bühlmann (2012). Given that some degrees of freedom exist within the authors' description, our implementation differs from that of other available approaches. The primary difference from the implementation by Augugliaro et al. (2023), aside from the variation in model selection, lies in the computation of the starting values for the covariance and means, as well as the penalty parameters. Our approach bears more similarities to Falk and Starr (2023), but still differs, for instance, in the definition of penalty parameters at the beginning. In this appendix, we provide a comprehensive account of our approach, which is applied with both glasso and atan regularization.

At the beginning, the data are standardized. To define the penalty parameters, we compute the covariance matrix (which corresponds to a correlation matrix due to the standardization) using pairwise deletion. The largest absolute value of the off-diagonal entries is multiplied by 1.001 and used as the largest penalty parameter. If no observations are available for a given pair of variables, a warning is issued, but the code continues by selecting from the remaining off-diagonal entries. The minimum penalty parameter is set as the largest penalty multiplied by 0.01. The penalty parameters are then distributed logarithmically between the minimum and maximum values, consistent with the behavior of the *qgraph* package.

The initial covariances and means for the EM algorithm are first calculated using listwise deletion, as described in Städler and Bühlmann (2012). If the resulting initial covariance matrix is not positive definite, it is adjusted to a near positive definite matrix using the *Matrix* package (Bates et al., 2023, Version 1.6-0). If listwise deletion is not feasible due to missing values on all observations, covariances between variables are set to zero. The inverse of the initial covariance matrix is then used as the initial precision matrix to compute conditional expectations in the E-step.

In the M-step, the covariance matrix calculated from the sufficient statistics is forced to be symmetric and positive definite using the *Matrix* package for regularization *via* glasso or atan. The precision matrix and means resulting from the M-step are compared to the values at the beginning of the E-step. If the difference for each parameter is smaller than a threshold of 0.00001, the algorithm terminates. Otherwise, the algorithm proceeds with the next E-step, continuing until the maximum number of iterations (1000) is reached.