Psychological Methods

Comparing Gaussian Graphical Models With the Posterior Predictive Distribution and Bayesian Model Selection

Donald R. Williams, Philippe Rast, Luis R. Pericchi, and Joris Mulder

Online First Publication, February 20, 2020. http://dx.doi.org/10.1037/met0000254

American Psychological Association

Psychological Methods

ISSN: 1082-989X

http://dx.doi.org/10.1037/met0000254

Comparing Gaussian Graphical Models With the Posterior Predictive Distribution and Bayesian Model Selection

Donald R. Williams and Philippe Rast
University of California, Davis

Luis R. Pericchi
University of Puerto Rico at Rio Piedras

Joris Mulder
Tilburg University

Abstract

Gaussian graphical models are commonly used to characterize conditional (in)dependence structures (i.e., partial correlation networks) of psychological constructs. Recently attention has shifted from estimating single networks to those from various subpopulations. The focus is primarily to detect differences or demonstrate replicability. We introduce two novel Bayesian methods for comparing networks that explicitly address these aims. The first is based on the posterior predictive distribution, with a symmetric version of Kullback-Leibler divergence as the discrepancy measure, that tests differences between two (or more) multivariate normal distributions. The second approach makes use of Bayesian model comparison, with the Bayes factor, and allows for gaining evidence for invariant network structures. This overcomes limitations of current approaches in the literature that use classical hypothesis testing, where it is only possible to determine whether groups are significantly different from each other. With simulation we show the posterior predictive method is approximately calibrated under the null hypothesis ($\alpha = .05$) and has more power to detect differences than alternative approaches. We then examine the necessary sample sizes for detecting invariant network structures with Bayesian hypothesis testing, in addition to how this is influenced by the choice of prior distribution. The methods are applied to posttraumatic stress disorder symptoms that were measured in 4 groups. We end by summarizing our major contribution, that is proposing 2 novel methods for comparing Gaussian graphical models (GGMs), which extends beyond the social-behavioral sciences. The methods have been implemented in the R package BGGM.

Translational Abstract

Gaussian graphical models are becoming popular in the social-behavioral sciences. Recently attention has shifted from estimating single networks to those from various subpopulations (e.g., males vs. females). We introduce Bayesian methodology for comparing networks estimated from any number of groups. The first approach is based on the posterior predictive distribution and it allows for determining whether networks are different from one another. This is ideal for testing the null hypothesis of group equality, say, in the context of testing for network replicability (or lack thereof). The second approach is based on Bayesian hypothesis testing and it allows for gaining evidence for network invariances or equality of partial correlations for any number of groups. This is ideal for focusing on specific aspects of the network such as individual partial correlations. In a series of simulations and illustrative examples we demonstrate the utility of the proposed methodology for comparing Gaussian graphical models. The methods have been implemented in the R package BGGM.

Keywords: Gaussian graphical model, posterior predictive distribution, Bayes factor, partial correlation

Donald R. Williams and Philippe Rast, Department of Psychology, University of California, Davis; Luis R. Pericchi, Department of Mathematics, University of Puerto Rico at Rio Piedras; Joris Mulder, Department of Methodology and Statistics, Tilburg University.

We thank Sacha Epskamp for suggestions that refined the presented methodology. Research reported in this publication was supported by five funding sources: (a) The National Academies of Sciences, Engineering, and Medicine FORD foundation pre-doctoral fellowship to Donald R. Williams; (b) The National Science Foundation Graduate Research Fellowship to Donald R. Williams; and (c) the National Institute On Aging of the National Institutes of Health under Award R01AG050720 to Philippe Rast; (d) Awards P20GM103475 from NIGMS and U54CA096297 from NCI of the NIH to Luis R. Pericchi; (e) an ERC Starting Grant (758791) to Joris Mulder. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Academies of Sciences, Engineering, and Medicine, the National Science Foundation, or the National Institutes of Health.

Correspondence concerning this article should be addressed to Donald R. Williams, Department of Psychology, University of California, Davis, 1 Shields Avenue, Davis, CA 95616. E-mail: drwwilliams@ucdavis.edu

The Gaussian graphical model (GGM) has become increasingly popular in the social-behavioral sciences (Epskamp & Fried, 2016). Traditional statistical approaches, for example the structural equation model (SEM) framework, conceptualize psychological constructs as arising from a common cause (i.e., latent variable; Cramer & Borsboom, 2015). Conversely, the primary motivation for GGMs is that *observed* variables are a dynamic, interacting system of relations (Epskamp, Waldorp, Mottus, & Borsboom, 2018). These effects are encoded in the inverse of the covariance matrix, in particular the off-diagonal elements, and correspond to the conditional (in)dependence structure of random variables (Dempster, 1972). When they are standardized and the sign reversed, this results in partial correlations that are pairwise relationships in which all other variables have been controlled for (Fisher, 1915; Yule, 1907). That is, when there is evidence for a nonzero effect, this indicates a direct association between two variables. The central objective, when estimating GGMs, is then to uncover the underlying psychological network that typically includes effects determined to be different than zero (but see Williams & Mulder, 2019a). Note that "network" is a generic term, that can apply to a variety of models (i.e., friendship; Marathe, Pan, & Apolloni, 2013), but here we are referring specifically to partial correlation networks. For the remainder of this work GGM and network are used interchangeably.

Not only are network models relatively new in the social-behavioral sciences, but there are few extensions that go beyond identifying the conditional (in)dependence structure. For example, only recently was an approach for confirmatory (Bayesian) hypothesis testing introduced in Williams and Mulder (2019a). While the improvement or development of novel estimation methods (e.g., penalized likelihood) is still an active area of research in the statistical literature (Fan, Liao, & Liu, 2016; Kuismin & Sillanpää, 2017), the focus is typically on increasing accuracy of point estimates or detection of nonzero partial correlations. This stands in contrast to SEM, where extensions are often introduced specifically for psychological applications (Preacher & Merkle, 2012). For example, a question of high interest is whether the same construct is being measured in different groups—that is, whether it is measurement invariant (van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015). This has resulted in a large body of literature (Muthén & Asparouhov, 2018), where establishing invariance is required for group comparisons (e.g., of factor scores; van de Schoot, Lugtig, & Hox, 2012), or testing the null hypothesis is the primary research question of interest (Verhagen & Fox, 2013; Verhagen, Levy, Millsap, & Fox, 2016).

Recently, the focus has shifted from estimating a network from one group, to comparing those estimated from different subpopulations (Fried et al., 2018). For example, group differences have been examined in depression networks (e.g., good vs. poor depression prognosis; Beard et al., 2016), as well as gender differences in hyper-sexuality networks (Werner, Štulhofer, Waldorp, & Jurin, 2018). These comparisons have sometimes been speculative, for example based visual inspection, or with a resampling approach that was recently introduced to psychology (van Borkulo et al., 2017). On the other hand, there has been an ongoing debate regarding the replicability of psychological networks (Forbes, Wright, Markon, & Krueger, 2019; Jones, Williams, & McNally, 2019). That is, group comparisons are not of primary interest but the focus is to replicate a given conditional (in)dependence structure in different groups. To our knowledge, all current methods for comparing GGMs rely on null hypothesis significance testing. This approach can only reject the null hypothesis of (typically) no effect but cannot provide evidence for the null hypothesis that networks are the same. Similar critiques also apply to classical measurement invariance testing procedures, for example as noted in Verhagen and Fox (2013) and Verhagen, Levy, Millsap, and Fox (2016), which partially motivates this work. In order to address these issues, we introduce novel Bayesian methods that allow for not only assessing group differences but also invariances. The latter can test the entire network or specific aspects (e.g., individual partial correlations).

This work is further motivated by additional limitations of existing methods. As noted, there is a resampling based approach, the network comparison test (NCT), that uses l_1-regularization to estimate the networks (van Borkulo et al., 2017). It is important to note that this approach does not *require* the use of l_1-regularization and it could be used with nonregularized approaches for estimating networks (Williams, Rhemtulla, Wysocki, & Rast, 2019). This method is not only computationally intensive, due to resampling and data driven model selection, but information is also lost with the chosen test statistics. For example, the test for *invariant* network structure is based on the maximum difference between two partial correlations in reference to a permutation distribution. As such, power to detect a difference depends completely on the magnitude of a single effect. We are aware of one additional classical (frequentist) approach for comparing GGMs that relies on desparisifying l_1-regularized point estimates (Belilovsky, Varoquaux, & Blaschko, 2016). In that approach, confidence intervals are constructed for testing differences between two partial correlations. This suffers from the same limitations as the NCT. In order to address these shortcomings, we propose a "global" approach that allows for testing the hypothesis of interest, that is, whether two networks were generated from different multivariate normal distributions—this is a critical assumption that underlies conditional independence coinciding with a partial correlation (Baba, Shibata, & Sibuya, 2004).

Together, the Bayesian methods introduced in this work were developed to overcome these limitations. First we introduce a "global" test that is based on a posterior predictive check. This test answers the question of whether there is some form of misfit in a model with equal networks across groups given the observed data. This is achieved by comparing the Kullback-Leibler divergence, which can be seen as a "distance" measure for distributions, between the expected networks of different groups, conditional on the observed data, with the Kullback-Leibler divergence from a model that assumes group equality. This considers all aspect of the network model, and essentially results in a predictive likelihood ratio that accounts for posterior uncertainty. Second we introduce a Bayesian model selection criterion that can answer which hypothesis out of a set of competing hypotheses best describes the observed data. This can be used to determine, for example, whether specific aspects of the networks are the same. We introduce "local" approaches for individual partial correlations. Here the differences are tested with the Bayes factor, which can provide relative evidence for the null hypothesis—that is, whether a specific partial correlation is the same across groups.

This work is organized as follows. We first introduce notation and nomenclature specific to GGMs. We then describe the pro-

posed "global" method based on posterior predictive loss functions, after which we examine numerical performance and then apply the methods to posttraumatic stress disorder symptoms. Next, Bayesian model selection is introduced for the "local" method, based on the recently developed matrix-F prior distribution. In a series of numerical experiments we examine sample size requirements for determining whether two GGMs are the same (in contrast to the predictive approach), in addition to detecting differences between two partial correlations with the Bayes factor. The extensive application integrates the predictive method and Bayesian model selection, for example by first testing whether groups are different and then asking specific questions about (possible) invariances in the estimated networks. We end by discussing limitation as well as future directions of the proposed methods.

The Gaussian Graphical Model

The Gaussian graphical model captures conditional relationships (Lauritzen, 1996) that are typically visualized to infer the underlying conditional (in)dependence structure (i.e., the "network"; Højsgaard, Edwards, & Lauritzen, 2012). The undirected graph is $G = (V, E)$, and includes a vertex set $V = \{1, \ldots, p\}$ as well as an edge set $E \subset V \times V$. Let $\mathbf{y} = (y_1, \ldots, y_p)^\top$ be a random vector indexed by the graphs vertices, of dimension p, that is assumed to follow a multivariate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and with a $p \times p$ positive definite covariance matrix $\boldsymbol{\Sigma}$. Without loss of information, the data is considered centered with mean vector 0. Denote the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. The graph is obtained from the off-diagonal elements $\theta_{ij} \in \boldsymbol{\Theta}_{ij}$. This is used to construct an adjacency matrix A that follows

$$A_{ij} = \begin{cases} 1, & \text{if } \theta_{ij} \neq 0, 1 \leq i < j \leq p \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

with $1 \leq i < j \leq p$ denoting the elements in the upper-triangular of the $p \times p$ matrix. Further, $(i, j) \in E$ when the variables i and j are *not* conditionally independent and set to zero otherwise. Note that the edges are partial correlations (ρ) determined to be nonzero. These are computed directly from the precision matrix as

$$\rho_{ij} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}, 1 \leq i < j \leq p. \tag{2}$$

These partial correlations are explicitly used for the Bayes factor based approaches, whereas the precision matrix is targeted for the posterior predictive method.

Posterior Predictive Distribution

The posterior predictive distribution plays a central role in Bayesian model checking (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019; Levy, Mislevy, & Sinharay, 2009; Sinharay & Stern, 2003). The idea is that generated data from the fitted model should look like the observed data $\mathbf{Y}$, which contains the response vector of person p on the p-th row for example. Hence, with n observation from each person, this results in a $n \times p$ data matrix $\mathbf{Y}$. In the case of a well-fitting model, the replicated data, herein referred to as $\mathbf{Y}^{rep}$, can be viewed as data that could have been observed (but were not) or as predictive data of future observations (Rubin, 1984). We adopt the latter perspective. This is summarized in Gelman, Meng, and Stern (1996):

> "as the data that would appear if the experiment that produced $\mathbf{Y}$ today were replicated tomorrow with the same model, $\mathcal{M}$, [and] the same (unknown) value of θ that produced $\mathbf{Y}$" (p. 737).

For our purposes, we extend "experiment" to the more general "data generating process." In the context of comparing GGMs, say, between two groups, the approach is to first estimate the GGM (i.e., $\boldsymbol{\Theta}$) conditional on all of the groups being equal. Then the posterior predictive distribution can be sampled from $\boldsymbol{\Theta}$. $\mathbf{Y}^{rep}$ then represents the data that we expect to observe in the future, assuming that the fitted model of group equality was the underlying data generating process. Of course, when comparing two groups, the same model is necessarily fit to both groups which allows for comparisons with the realized predictive distribution under group equality. Given that the predictive distribution can be obtained from any number of groups, this approach seamlessly expands to situations where we wish to compare more than two groups. This is also a novel aspect of this work, in that the permutation based method is specifically for two groups (van Borkulo et al., 2017).

The posterior predictive distribution, for the purpose of model checking, is not without limitations (Robins, van der Vaart, & Ventura, 2000). For example, it has been criticized for double use of data (Dahl, Gasemyr, & Natvig, 2007) and that it is overly conservative (i.e., low "power" to detect misfit; Meng, 1994). The latter is attributed to the fact that posterior predictive p values are not uniform under the null-hypothesis (Gelman, 2013; van Kollenburg, Mulder, & Vermunt, 2017). Although there have been proposals to achieve calibration (Bayarri & Berger, 2000; Hjort, Dahl, & Steinbakk, 2006; van Kollenburg et al., 2017), our approach does not aim to be calibrated in the frequentist sense. Of course, posterior predictive model checking does share similarities with classical methods (Gelman, 2013)—for example, tail area probabilities are computed from repeatedly sampling an assumed model and that it is not possible to gain evidence for the null hypothesis. This also applies to this method, in that only group differences can be assessed (but see Bayesian Hypothesis Testing section). Furthermore, we are not model checking in the typical sense, but explicitly testing whether two or more precision matrices were generated from different multivariate normal distributions.

Method Description

We first introduce the customary notation, for the univariate case, which serves as the foundation for our method. The observed data is denoted by $\mathbf{Y}$, a fitted model is denoted by $\mathcal{M}$, and the parameters to be estimated is $\boldsymbol{\theta}$, with prior distribution $p(\boldsymbol{\theta})$. The posterior predictive distribution is then

$$p(\mathbf{Y}^{rep} \mid \mathcal{M}, \mathbf{y}) = \int p(\mathbf{Y}^{rep} \mid \mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}, \mathbf{Y}) d\boldsymbol{\theta}. \tag{3}$$

Note that $\mathbf{Y}^{rep}$ can be compared visually with $\boldsymbol{y}$, but for computing posterior predictive p values, herein referred to as p values, a test-statistic T is needed which is a function of an observed or replicated data set. This allows for comparing $T(\mathbf{Y}_{rep})$ with the observed $T(\mathbf{Y})$—for example,

$$p\text{-value} = p[T(\mathbf{Y}^{rep}) > T(\mathbf{y}) \mid \mathcal{M}, \mathbf{Y}]. \quad (4)$$

This is the probability that $T(\mathbf{Y}^{rep})$ is greater than $T(\mathbf{Y})$, conditional on $\mathcal{M}$ and $\mathbf{Y}$. This is computed as the proportion of $T(\mathbf{Y}^{rep})$ that exceed $T(\mathbf{Y})$. Note that the replicated data sets are obtained from drawing samples from the posterior distribution of $\boldsymbol{\Theta}$. This is further clarified below.

We now extend this notation to multivariate data from possibly multiple groups. We first assume that each group $g \in \{1, \ldots, G\}$ is a realization from the same multivariate normal distribution—that is, the null model

$$\mathcal{M}_0 : \boldsymbol{\Theta}_1 = \ldots = \boldsymbol{\Theta}_G. \quad (5)$$

The posterior for the common precision matrix $\boldsymbol{\Theta}(= \boldsymbol{\Theta}_1 = \ldots = \boldsymbol{\Theta}_G)$, given the observed data, can be written as $p(\boldsymbol{\Theta} \mid \mathbf{Y}_1^{obs}, \ldots, \mathbf{Y}_G^{obs}, \mathcal{M}_0)$. Under $\mathcal{M}_0$, a posterior draw(s) for $\boldsymbol{\Theta}^{(s)}$ is in fact a posterior draw for the precision matrix in all groups, that is, $\boldsymbol{\Theta}^{(s)} = \boldsymbol{\Theta}_1^{(s)} = \ldots = \boldsymbol{\Theta}_G^{(s)}$. To simplify computing the posterior distribution we use the improper Jeffreys prior. This allows for sampling directly from a Wishart distribution—for example,

$$\boldsymbol{\Theta}(= \boldsymbol{\Theta}_1 = \ldots = \boldsymbol{\Theta}_G) \sim W(n - 1, \boldsymbol{S}^{-1}), \quad (6)$$

where n is the sample size (of all groups combined) and $\boldsymbol{S}$ denotes the scatter matrix $\mathbf{Y}'\,\mathbf{Y}$ (for all groups as well; Gelman et al., 2014). Note that (6) is the distribution conditional on Y. Next we generate a replicated data set given these precision matrices—for example,

$$\begin{aligned} \boldsymbol{\Theta}_1^{(s)} &\rightarrow \mathbf{Y}_1^{rep^{(s)}} \\ &\vdots \\ \boldsymbol{\Theta}_G^{(s)} &\rightarrow \mathbf{Y}_G^{rep^{(s)}}. \end{aligned} \quad (7)$$

Note that, in the case of unequal group sizes, these replicated data sets are generated with the observed group sizes. Now the posterior expectation of a precision matrix for group g given $\mathbf{Y}_g^{rep}$ can be approximated as

$$E\{\boldsymbol{\Theta}_g^{rep} \mid \mathbf{Y}_g^{rep}\} = (n_g - 1)\big(\mathbf{Y}_g^{rep\prime}\mathbf{Y}_g^{rep}\big)^{-1}. \quad (8)$$

This approximation is the inverse of unbiased estimate of the sample based covariance matrix, which will coincide (approximately) with the posterior expectation in the case of an improper prior distribution (6).

In review it was pointed out that focusing on $\boldsymbol{\Theta}$ is not ideal, because it includes the diagonal elements that are not important for network *inference*. A test using (8) could result in detecting a difference that is attributable to the variance. However, two groups could have the same underlying partial correlation network. To remove the effects of Θ_{ii}, we follow the approach described in Padmanabhan, White, Zhou, and O'Connell (2016) and use the normalized precision matrix. This is accomplished with the following parameterization

$$\boldsymbol{\Theta} = \mathbf{DRD}, \quad (9)$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sqrt{\boldsymbol{\Theta}}_{ii}$ and $\mathbf{R}$ has $r_{ij} = \Theta_{ij}/\sqrt{\Theta_{ii}\Theta_{jj}}$ on the off-diagonals and one on the diagonal. This is similar to the parameterization described in Epskamp, Rhemtulla, and Borsboom (2017). In our formulation, this effectively separates out the diagonal elements of $\boldsymbol{\Theta}$. Note $\mathbf{R}$ is *not* the partial correlation—that would require reversing the direction ($\pm$) of r_{ij}. However, we found that reversing the direction can result in ill-conditioned matrices that does not allow for computing the chosen test statistic. Hence we use of the normalized precision matrix $\mathbf{R}$ for the predictive check.

Network predictive check. This approach is meant to parallel the network structure *invariance* test in van Borkulo et al. (2017). Of note, while the name implies a test for the null hypothesis (i.e., no-difference), it only can determine differences. Because this also applies to our approach, we avoid the word *invariance* until later on (see Bayesian Hypothesis Testing section). In van Borkulo et al. (2017) the maximum difference between two edges, in reference to a permutation distribution for two groups, was taken to indicate whether the network structures differed. Our aim is the directly assess whether two or more GGMs, while accounting for posterior uncertainty, were generated from different multivariate normal distributions. For the test-statistic we thus use a version of Kullback-Leibler divergence (KLD), which is also known as entropy loss (Kuismin & Sillanpää, 2017), is proportional (i.e., by $\frac{1}{2}$) to Stein's loss for covariance matrices (e.g., Equation 72 in James & Stein, 1961), and is the log likelihood ratio between two distributions (Eguchi & Copas, 2006). Note that KLD has several motivations, for example maximizing the likelihood is equivalent to minimizing KLD between two distributions (Grewal, 2011). Further, in Bayesian contexts, it has been used for selecting models (Goutis, 1998; Piironen & Vehtari, 2017) and prior distributions (Bernardo, 2005), variational inference (Blei, Kucukelbir, & McAuliffe, 2017), and is known to be minimized by the Bayes factor (when used for model selection) in so-called $\mathcal{M}$-open settings (Bernardo & Smith, 2001; Yao, Vehtari, Simpson, & Gelman, 2018).

These uses have one common theme—that is, assessing the distance between distributions. However, KLD is not a true distance measure because it is asymmetric. As such, we use Jensen-Shannon divergence (JSD) which symmetrizes KLD (Nielsen, 2010). For two randomly selected groups, the test-statistic is then

$$T = \text{JSD}(E\{\mathbf{R}_{g_1} \mid \mathbf{Y}_{g_1}\}, E\{\mathbf{R}_{g_2} \mid \mathbf{Y}_{g_2}\}), \quad (10)$$

which is the average KLD in both directions—that is,

$$\begin{aligned} \text{JSD} = \frac{1}{2}[&\text{KLD}(E\{\mathbf{R}_{g_1} \mid \mathbf{Y}_{g_1}\}, E\{\mathbf{R}_{g_2} \mid \mathbf{Y}_{g_2}\}) \\ &+ \text{KLD}(E\{\mathbf{R}_{g_2} \mid \mathbf{Y}_{g_2}\}, E\{\mathbf{R}_{g_1} \mid \mathbf{Y}_{g_1}\})]. \end{aligned} \quad (11)$$

For a multivariate normal distribution KLD is defined as

$$\text{KLD}(\mathbf{R}_{g_1} \parallel \mathbf{R}_{g_2}) = \frac{1}{2}\Big[\text{tr}\big(\mathbf{R}_{g_1}^{-1}\mathbf{R}_{g_2}\big) - \log\big(\mid \mathbf{R}_{g_1}^{-1}\mathbf{R}_{g_2} \mid\big) - p\Big], \quad (12)$$

where p is the number of variables. Note that inverting $\mathbf{R}_{g_1}^{-1}$ results in the covariance matrix $\mathbf{R}_{g_1}$ and E[.] has been removed to simplify (12). Repeating this process for each posterior sample produces the predictive distribution of JSD. To be clear, this distribution can be thought of as the amount of divergence (or relative entropy) we would expect to see assuming that the null model of group equality were *true*. This serves as the reference distribution, from which the predictive p value is computed as

$$p = \frac{1}{S}\sum_{s=1}^{S} I\left(T\left(\mathbf{Y}_1^{obs}, \ldots, \mathbf{Y}_G^{obs}\right) < T\left(\mathbf{Y}_1^{rep^{(s)}}, \ldots, \mathbf{Y}_G^{rep^{(s)}}\right)\right), \tag{13}$$

where $I(\cdot)$ is the indicator function. A decision rule is required for determining whether the two GGMs are "significantly" different from each other (i.e., p value $\leq \alpha$). This leaves open the choice of α which can either be determined based on subjective grounds or with guidance from the present numerical experiments (or a combination of both).

Procedure. To summarize, this method follows these steps:

1. Estimate $p(\boldsymbol{\Theta} \mid \mathbf{Y}_1^{obs}, \ldots, \mathbf{Y}_G^{obs}, M_0)$ with (6).
2. For each posterior sample (s)
   (a) $\Theta_g^{(s)} \rightarrow \mathbf{Y}_g^{rep^{(s)}}$, for $g \in \{1, \ldots, G\}$.
   (b) Compute $\mathbf{R}_g^{rep^{(s)}}$
   - $\mathbf{R}_g^{rep^{(s)}} = \boldsymbol{d}_g^{rep^{(s)}}\Theta_g^{rep^{(s)}}\boldsymbol{d}_g^{rep^{(s)}}$, where $\boldsymbol{d}_g^{rep^{(s)}}$ is a diagonal matrix with $d_{ii}^{rep^{(s)}} = 1\sqrt{\Theta_{ii}^{rep^{(s)}}}$.
   - $\Theta_g^{rep^{(s)}} = (n-1)\boldsymbol{S}^{-1}$, where $\boldsymbol{S} = \mathbf{Y}_g^{rep^{(s)\prime}}\mathbf{Y}_g^{rep^{(s)}}$
   (c) Compute the predictive "distance":
   JSD $(E\{\mathbf{R}_{g_1}^{rep} \mid \mathbf{Y}_{g_1}^{rep}\}, E\{\mathbf{R}_{g_2}^{rep} \mid \mathbf{Y}_{g_2}^{rep}\})$.
3. Compute the observed "distance":
   JSD $(E\{\mathbf{R}_{g_1}^{obs} \mid \mathbf{Y}_{g_1}^{obs}\}, E\{\mathbf{R}_{g_2}^{obs} \mid \mathbf{Y}_{g_2}^{obs}\})$.
4. Compute the posterior predictive p value with (4).

Note that g_1 and g_2 were used to keep the notation manageable. This procedure can apply to any number of groups (with possibly unequal means and standard deviations because we standardize the data).

At this point, it is worth emphasizing that the predictive method is not restricted to (symmetric) KL-divergence—the method is general. For example, the package **NCT** uses the maximum partial correlation difference between two networks or the "weighted absolute sum of all edges in the network" (van Borkulo et al., 2017, p. 8). These could also be used as test statistics in the predictive method, although we think it is important to consider other possibilities for comparing networks. This is discussed further in the Future Directions section.

Nodewise predictive check. The network approach is "global," in that all aspects of the *normalized* precision matrices are being tested. It is also important to consider more targeted comparisons, particularly in the event $\mathcal{M}_0$ is rejected. We thus extend the method to consider predictive KL-divergence of each node in the network. This is a result of the direct correspondence between the elements of Θ and regression coefficients (Kwan, 2014; Stephens, 1998)—for example,

$$\theta_{ij} = -\frac{\beta_{ij}}{\sigma_j^2} \quad \text{and} \quad \theta_{jj} = \frac{1}{\sigma_j^2}, i \neq j, \tag{14}$$

Here j denotes the respective column of the $p \times p$ matrix and σ_j^2 is the residual variance from the jth regression model, where the jth column is predicted by the remaining $(p-1)$ variables. Further details can be found in Williams (2018). This relationship allows for directly building upon the previously described method by estimating the respective regression coefficients from $\Theta_g^{(s)} \rightarrow \mathbf{Y}_g^{rep^{(s)}}$ for $g \in \{1, \ldots, G\}$. Then KL-divergence is computed based on the predictive distribution as

$$\hat{\mathbf{y}}_{g,j}^{rep^{(s)}} = \mathbf{Y}_{g,-j}^{rep^{(s)}}\beta_{g,j}^{(s)}, \tag{15}$$

where "$-j$" denotes removal of that specific column, as it is the outcome variable, $\beta_{g,j}^{rep^{(s)}}$ is a $(p-1)$ vector of estimated regression coefficients (with least squares), and $\hat{\mathbf{y}}_{g,j}^{rep^{(s)}}$ is the predicted values for the jth variable. Because the data was scaled in advance, this simplifies the calculation of KL-divergence by only having to consider the variance of $\hat{\mathbf{y}}_{g,j}^{rep^{(s)}}$—for example,

$$\mathrm{KLD}\left(\hat{\mathbf{y}}_{g_1,j}^{rep^{(s)}} \| \hat{\mathbf{y}}_{g_2,j}^{rep^{(s)}}\right) = \log\frac{\sigma_{g_2,j}^{rep^{(s)}}}{\sigma_{g_1,j}^{rep^{(s)}}} + \frac{\sigma_{g_1,j}^{2rep^{(s)}}}{2\sigma_{g_2,j}^{2rep^{(s)}}} - 0.5. \tag{16}$$

$\sigma_{g,j}^{2rep^{(s)}}$ is the variance of the predictive distribution for each replicated data set and j denotes the node under consideration. This can similarly be symmetrized, by taking the average of both directions, which results in Jensen-Shannon divergence. Furthermore, the p value is computed as in Equation 4 but with respect to each variable in the network. This allows for testing whether each node, for any number of groups, is different from one another according to the predictive distribution and chosen α level. Note that the following experiments only look at the network approach (see Network Predictive Check section), but the null distribution, assuming group equality, was similar for both approaches.

Numerical Performance

Null distribution. Posterior predictive p values, defined in Equation 4, are not necessarily calibrated in the frequentist sense. That is, under the null hypothesis classical p values $\in [0, 1]$ are equally likely which results in a uniform distribution. This is not necessarily the case for the present p values. We thus examined the null-distribution for Jensen-Shannon divergence (10), where the null hypothesis of group equality was true. In particular, we set $G = 2$ and $n \in \{250, 500, \text{and } 1{,}000\}$. We also examined unequal group sizes by reducing the sample size of one group by 50%—for example, $n_{g_1} = 250$ and $n_{g_2} = 125$. All of the simulations used correlations matrices from Fried et al. (2018), which included posttraumatic stress symptoms from four groups. This decision was made because we wanted the population values and level of sparsity (i.e., the proportion of zeroes) to be representative of a common psychological application in the network literature. For this simulation in particular, we used the largest sample size ($N = 956$ and $p = 16$). We first converted the correlation matrix to the partial correlation matrix, set values less than 0.05 to zero (Epskamp, 2016), then treated this as the true network structure for each group. Each condition was repeated for 1,000 simulation trials.

We first plotted representative predictive distributions (Figure 1; panel A). The corresponding observed divergence is also included (the black dots), each of which was not surprising such that the null hypothesis, that is, $\mathcal{M}_0$, would not be rejected ($\alpha = .05$). Note that this is an explicitly one-sided test in that we are only concerned with more divergence under the fitted $\mathcal{M}_0$ than the observed

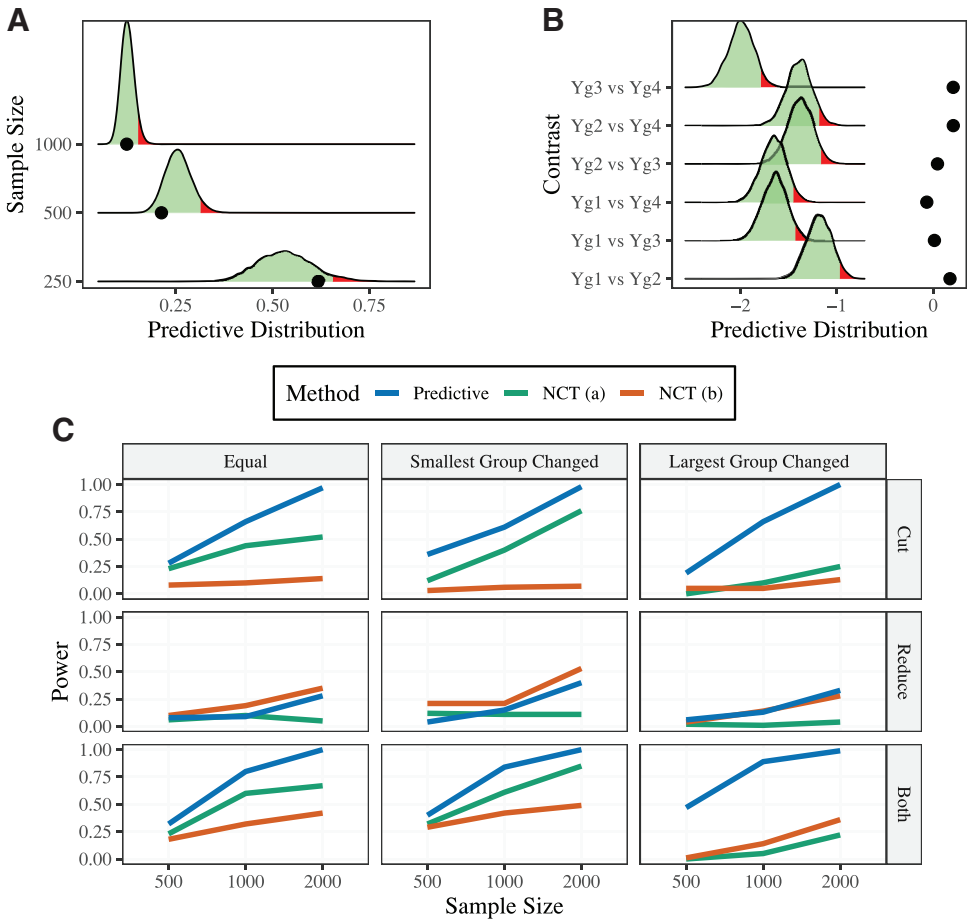

Figure 1. (A) Representative predictive distributions of JSD (symmetric KL-divergence) under the null hypothesis (M_0). The observed error is denoted with the black points and the red area is the critical region (α = .05). The posterior predictive p value (4) is the density to the right of the observed error. (B) Predictive distributions for pairwise comparisons between four groups (Posterior Predictive Distribution). The observed error is denoted with the black points. The density greater than the observed is the p value, which in this case, is 0 for all comparisons. (C) Simulation results (Detecting Differences). The x-axis denotes the total sample size of both groups combined. Unequal groups were divided: 60% and 40% of the total sample size. NCT (a): global strength. NCT (b): maximum difference. Cut: edges smaller than 0.075 were set to zero. Reduce: the largest edge was reduced by 25% (creating a difference greater than 0.10). Both: edges were cut and the largest was reduced. NCT = network comparison test. See the online article for the color version of this figure.

divergence. This visualization shows the effect of sample size on the predictive distribution, in that the expected divergence, assuming group equality, reduced with larger sample sizes. Note that this behavior is also typically observed for the sampling distribution in classical significance tests such as in the classical t test. Furthermore, as seen in Table 1, it appears that the error rate is close to the nominal level of 0.05. Of course, from a Bayesian perspective the goal is not necessarily to be calibrated in the frequentist since, so long as it is still possible to reliably detect differences. Although not discussed here, the error rates were similar when considering more than two groups.

Detecting differences. Here we examine power for detecting differences between two GGMs. Because our method is different than the NCT, it was not entirely clear how best to compare their performances. For example, while we could have implemented an approach that tests the maximum difference based on the predictive distribution, this would not take full advantage of KL-divergence that is the expected log likelihood ratio (Eguchi & Copas, 2006). We thus followed a similar approach as van Borkulo et al. (2017), in that we manipulated the strongest edge, reduced some edges to zero, and also a combination of both. First, the same correlation matrix (see Null Distribution section) was converted to the partial correlation matrix, and then values less than 0.05 were set to zero. This served as the baseline, and for the *subtle* manipulations, we either reduced the largest edge by 25%, set additional values to zero (i.e., also those less than 0.075), or a combination of both. These network structures are provided in Appendix A (Figure A1). The total sample size was fixed to 500, 1,000, and 2,000. For the unequal conditions the largest sample size was 60% of the total—for example, n_{g_1} = 1,200 and n_{g_2} = 800. We further

Table 1
Error Rate for Rejecting the (True) Null Model

Measure	*n*	Sample composition	Error rate	MCE
JSD	250	Equal	0.052	0.002
		Unequal	0.043	0.001
	500	Equal	0.047	0.001
		Unequal	0.048	0.001
	1,000	Equal	0.041	0.001
		Unequal	0.046	0.001

Note. JSD = Jensen-Shannon divergence. $\alpha = .05$. MCE = Monte Carlo error rounded to the third decimal place. The provided sample size corresponds to the largest group for the unequal conditions (the smaller group was half that size).

manipulated which group, that is the largest or smallest, had the altered network structure. We used the default settings in the NCT package, and the *p* values for both network "invariance" and global strength (which sums the absolute errors between partial correlations matrices) were collected. The alpha level was set to 0.05 and each condition was repeated for 100 simulation trials.

Both methods require repeated sampling. The NCT performs data-driven model selection for each permutation sample, whereas our method first samples from the posterior and then from the predictive distribution. We thus looked at the speed of each method per 1,000 iterations. The results are provide in the Appendix B. The predictive approach was faster than the permutation based NCT. This highlights the computational efficiency of our method. Note that the predictive approach did require more time with larger sample sizes, whereas sample sizes did not seem to matter for the NCT. Still, that the NCT took more than eight times longer for the largest sample size ($n = 1{,}000$) indicates computational feasibility is not an issue with the predictive method.

The simulation results are provided in Figure 1 (panel C). Because our method considers the entire precision matrix, we compared it to both NCT approaches for all conditions (although each is for a specific test statistic). The predictive method not only had competitive performance, but for almost all conditions, the power was higher than both NCT approaches. In particular, with different conditional independence structures (Figure 1; "Cut" and "Both"), the predictive method had much higher power to detect the differences. Note that cutting of edges effectively created differences of 0.075 or less, which would take a considerable sample size to detect for the maximum difference NCT. This is because it focuses on only one difference, whereas our use of JSD can be understood as a multivariate log likelihood ratio that also incorporates posterior uncertainty. Note that the maximum difference NCT did have the most power when *only* the largest edge in the network was reduced. The power was also low, for all methods, when the maximum edge was reduced but the network (e.g., the conditional independence structure) otherwise stayed the same. However, with *subtle* differences in both the conditional independence structure and a small difference in the strongest edge, the predictive method excelled by capturing all aspects of the normalized precision matrix. Further, as shown in the panel "Largest Group Changes," the predictive approach was less sensitive to unequal sample sizes. It is important to emphasize that these changes to the networks were small, as seen in Figure A1, which indicates the predictive method has high power while also maintaining the nominal α level (see Table 1).

Bayesian Hypothesis Testing

Although the predictive method did well at detecting differences between networks structures, it cannot provide evidence for a null model that assumes that certain edges have equal strengths across groups. Further, the predictive approach is essentially an omnibus test: it does not provide specific information about the differences between groups. We thus compliment the "global" predictive method with a "local" Bayes factor test, that allows for focusing on particular aspects of the network. The key difference is that the following does not attempt to reject the null model (i.e., $\mathcal{M}_0$ that groups are the same), but compares models to assess the relative evidence in the data between competing hypotheses. For example we could quantify the evidence in favor of H_0: the groups are (exactly) the same against, H_1: the groups are not (exactly) the same, or we could test differences between specific partial correlations. In contrast to the predictive approach, that used an improper Jeffreys prior (6), the Bayes factor test requires proper prior distributions for all parameters that are tested (e.g., Jeffreys, 1961).

A Matrix-*F* Distributed Conjugate Prior

The matrix-*F* was recently proposed as a flexible alternative to the inverse Wishart and Wishart prior for covariance and precision matrices, respectively (Mulder & Pericchi, 2018). To our knowledge this prior has only been employed once in the context of GGMs (Williams & Mulder, 2019a). We specify an encompassing matrix-*F* prior distribution for the precision matrix,

$$\mathbf{\Theta} \sim F(\nu, \delta, \mathbf{B}), \tag{17}$$

where $\nu > p - 1$ and $\delta > 0$ are the first and second degrees of freedom, which control the behavior near the origin and in the tails, respectively, and $\mathbf{B}$ is a positive definite scale matrix. For completeness the prior density of the matrix-*F* prior is given in Appendix A and further details about the encompassing prior approach for hypothesis testing can be found in Klugkist, Kato, and Hoijtink (2005). The matrix-*F* prior can be written as a scale mixture of Wishart distributions with an inverse Wishart mixture distribution, i.e.,

$$\begin{aligned} \mathbf{\Theta} \mid \mathbf{\Psi} &\sim W(\nu, \mathbf{\Psi}) \\ \mathbf{\Psi} &\sim IW(\delta + p - 1, \mathbf{B}). \end{aligned} \tag{18}$$

Because the Wishart prior is conjugate, it follows that the matrix-*F* prior is conditionally conjugate. That is, the conditional posterior of $\mathbf{\Theta}$ given $\mathbf{\Psi}$ has a Wishart distribution and the conditional posterior of $\mathbf{\Psi}$ given $\mathbf{\Theta}$ has an inverse Wishart distribution (Appendix A). This makes the matrix-*F* prior computationally feasible for GGMs, in that the posterior can be obtained with a Gibbs sampler (Appendix A).

The hypotheses of interest are not directly formulated on $\mathbf{\Theta}$, but on the partial correlations ρ given in Equation 2. To understand the *implied* marginal prior for ρ_{ij}, consider the fact that the matrix-*F* prior can be written as a scale mixture of inverse Wishart distributions with a Wishart mixture distribution—for example,

$$\begin{aligned} \boldsymbol{\Theta} \mid \boldsymbol{\Phi} &\sim IW(\delta + p - 1, \boldsymbol{\Phi}) \\ \boldsymbol{\Phi} &\sim W(\nu, \mathbf{B}). \end{aligned} \tag{19}$$

Furthermore, due to Barnard, McCulloch, and Meng (2000) it is known that a covariance matrix having an inverse Wishart prior distribution with an identity scale matrix, that is, $IW(\nu, \mathbf{I}_p)$, results in marginal priors for the bivariate correlations having *beta* $\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ distributions in the interval $(-1, 1)$. Consequently, if a precision matrix has an inverse Wishart prior distribution, that is, $\boldsymbol{\Theta} \sim IW(\delta + p - 1, \mathbf{I}_p)$, the partial correlations then follow a $beta\left(\frac{\delta}{2}, \frac{\delta}{2}\right)$ distribution in the interval $(-1, 1)$, which is invariant to the dimension of the network p. We therefore set $\mathbf{B} = \epsilon\mathbf{I}_p$ and $\nu = \epsilon^{-1}$, for a small value for ϵ (e.g., 0.001), so that $\boldsymbol{\Phi} \approx \mathbf{I}_p$ and $\boldsymbol{\Theta}$ is approximately distributed as $IW(\delta + p - 1, \mathbf{I}_p)$.

In sum, the prior for the precision matrix and the implied marginal prior for the partial correlations are specified as

$$\begin{aligned} \boldsymbol{\Theta} &\sim F(\epsilon^{-1}, \delta, \epsilon\mathbf{I}_p) \\ \rho_{ij} &\sim beta\left(\frac{\delta}{2}, \frac{\delta}{2}\right) \text{ on } (-1, 1), \end{aligned} \tag{20}$$

for $i \neq j = 1, \ldots, p$, respectively. The prior hyperparameter δ can be chosen such that the prior standard deviation corresponds with the expected deviation from zero in the case of a partial correlation would be unequal to zero. Because the prior standard equals $s_\rho = 1/\sqrt{\delta + 1}$, which is the standard deviation of a beta distribution, one can set the hyperparameter equal to $\delta = (s_\rho^2)^{-1} - 1$ by plugging in the anticipated deviation from zero of the partial correlations for s_ρ.

Pairwise Hypothesis Testing

In this section we present a Bayes factor for testing whether partial correlations between variable i and j are equal across groups,

$$H_{0,ij}: \rho_{ij,1} = \ldots = \rho_{ij,G} \text{ vs. } H_{1,ij}: \text{“not } H_{0,ij}\text{”},$$

under the alternative hypothesis the partial correlations of at least two groups are unequal. The constraints under the null hypothesis can compactly be formulated as $\mathbf{R}_{ij}\boldsymbol{\rho} = \mathbf{0}$, where $\mathbf{R}_{ij}$ is a matrix with coefficients capturing the equality constraints. The hypothesis test can then be written as $H_{0,ij} : \mathbf{R}_{ij}\boldsymbol{\rho} = \mathbf{0}$ versus $H_{1,ij} : \mathbf{R}_{ij}\boldsymbol{\rho} \neq \mathbf{0}$. For example, in the simple case of a network with three variables and two groups, the hypothesis can be written as $H_{0,ij} : \rho_{12,1} = \rho_{12,2}$, the parameter vector as $(\rho_{12,1}, \rho_{13,1}, \rho_{23,1}, \rho_{12,2}, \rho_{13,2}, \rho_{23,2})$, and the coefficients matrix as $\mathbf{R}_{12} = [1\,0\,0\;\; -1\,0\,0]$.

When testing a precise hypothesis with certain equality constraints on the parameters of interest, it is well-known that the prior for the free parameters under the alternative should be carefully chosen based on the anticipated effects (Bartlett, 1957; Jeffreys, 1961; Lindley, 1957). If the prior is unrealistically vague, it places too much probability mass at unrealistic values of the parameters, resulting in an overestimation of the evidence for the null when observing moderately sized effects. On the other hand if the prior is too informative by placing too much probability mass near the origin, it becomes difficult to distinguish between the null and the alternative hypothesis when quantifying the relative evidence in the data between the hypotheses. An example of this, for GGMs in particular, is provided in Williams and Mulder (2019a) Table C.3.

Due to the importance of the prior standard deviation under the alternative, the flexibility of the matrix-F prior becomes particularly useful by choosing δ such that the prior reflects the anticipated magnitude of the effects before observing the data. This can be done regardless of the network size p. Figure 2 displays the implied prior for $\rho_{ij,g}$ (left panel) as well as the implied prior for the difference of partial correlations between two groups $\rho_{ij,g} - \rho_{ij,g-1}$, for $\delta = 2$, 15, and 99, corresponding to prior standard

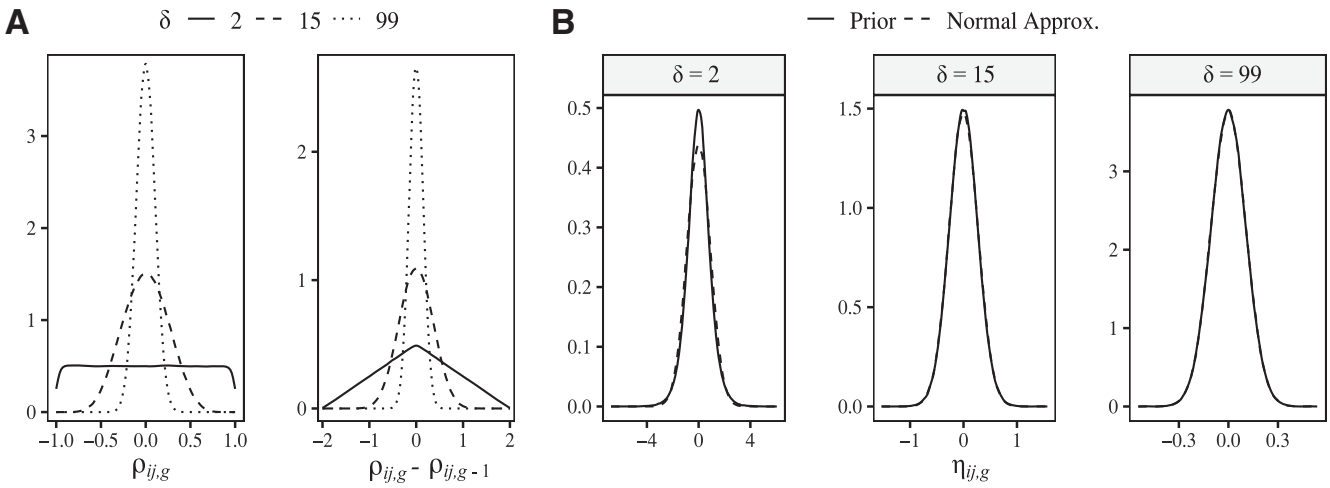

Figure 2. (A) Marginal prior distributions. Left panel: Marginal prior for the partial correlation between variables i and j in group g for a prior hyperparameter of $\delta = 2$ (solid line), 15 (dashed line), and 99 (dotted line), which corresponds to prior standard deviations of .58, .25, and .10, respectively. Right panel: Marginal prior for the difference between the partial correlation between variables i and j (in two different groups) and based on the same prior hyperparameters. (B) Prior of Fisher transformed partial correlation $\eta_{ij,g} = F(\rho_{ij,g})$ (solid line) and corresponding normal approximation (dashed line) for $\delta = 2$ (left panel), $\delta = 15$ (middle panel), and $\delta = 99$ (right panel).

deviations of .58, .25, and .10 for $\rho_{ij,g}$, respectively. Note that $\rho_{ij,g} - \rho_{ij,g-1}$ equals 0 under the above null hypothesis.

Now that the prior is specified, we can quantify the relative evidence between the hypotheses via the Bayes factor using the Savage-Dickey density ratio (Dickey, 1971; Mulder, Hoijtink, & Klugkist, 2010; Wetzels, Grasman, & Wagenmakers, 2010), which is defined as the ratio of the posterior and prior density evaluated at the null value under an unconstrained model—for example,

$$B_{01,ij} = \frac{p_u(\mathbf{R}_{ij}\boldsymbol{\rho} = \mathbf{0} \mid \mathbf{Y})}{p_u(\mathbf{R}_{ij}\boldsymbol{\rho} = \mathbf{0})}, \tag{21}$$

where p_u in the numerator and denominator denote the unconstrained posterior and prior density. The posterior and prior density in the numerator and denominator in Equation 21 do not have analytic expressions. In the simple case of $G = 2$ groups, we can get an accurate estimate of the posterior and prior density of $\rho_{ij,2} - \rho_{ij,1}$ at 0. This can be accomplished by first obtaining posterior and prior draws for the partial correlations, subtracting those to get the posterior and prior draws for the difference between partial correlations, and then finding the posterior and prior density (of the difference) evaluated at zero. This can be computed with the `density` or `logspline` functions in `R` (Deng & Wickham, 2011).

In the general case of more than two groups, the respective multivariate posterior and prior densities cannot be estimated using those R-functions. In that case we get an accurate and computationally feasible estimate of the posterior and prior density by following these steps:

1. Get S prior and posterior draws for $\boldsymbol{\rho}$ by sampling from the matrix-F prior and by using the Gibbs sampler (Appendix A).

2. Apply a Fisher transformation to the drawn partial correlations, i.e.,

$$\eta_{ij,g}^{(s)} = F\left(\rho_{ij,g}^{(s)}\right) = \frac{1}{2}\log\left(\frac{1+\rho_{ij,g}^{(s)}}{1-\rho_{ij,g}^{(s)}}\right), \quad \text{for} \quad s = 1, \ldots, S. \tag{22}$$

3. Compute the Fisher transformed differences via $\boldsymbol{\xi}^{(s)} = \mathbf{R}_{ij}\boldsymbol{\eta}^{(s)}$, for draws $s = 1, \ldots, S$. These transformed parameters are approximately normally distributed in the prior and posterior as shown below. Note that in terms of these transformed parameters, the hypothesis test can be written as $H_{0,ij} : \boldsymbol{\xi} = \mathbf{0}$ versus $H_{1,ij} : \boldsymbol{\xi} \neq \mathbf{0}$.

4. Estimate the posterior mean vector $\boldsymbol{\mu}_{\xi,N}$ and covariance matrix $\boldsymbol{\Psi}_{\xi,N}$, and the prior covariance matrix $\boldsymbol{\Psi}_{\xi,0}$ from their respective posterior and prior samples. Note that the prior mean vector equals $\mathbf{0}$.

5. Estimate the Bayes factor using

$$B_{01,ij} \approx \frac{N(\mathbf{0};\, \boldsymbol{\mu}_{\xi,N}, \boldsymbol{\Psi}_{\xi,N})}{N(\mathbf{0};\, \mathbf{0}, \boldsymbol{\Psi}_{\xi,0})}, \tag{23}$$

where $N(\mathbf{0}; \boldsymbol{\mu}, \boldsymbol{\Psi})$ denotes a multivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Psi}$ evaluated at $\mathbf{0}$.

The approximate normality of the posterior in Step 4 can be understood from the well-known fact that the sampling distribution (i.e., the likelihood) of a Fisher transformed correlation is approximately normally distributed (Fisher, 1915, 1921). Furthermore, the prior of a Fisher transformed partial correlation, $\eta_{ij,g}$, is also approximately normally distributed, as can be seen from Figure 2 for different values for $\delta = 2$, 15, and 99. Importantly, for small values of δ the approximation is slightly off near the origin, whereas for larger values of δ the approximation is very accurate. Note that typically one would not set a very small value for δ, as to avoid placing too much prior probability mass on unrealistically large effects. Consequently, combining an approximately normal prior with an approximately normal likelihood results in an approximately normal posterior for $\eta_{ij,g}$ (Mulder, 2016). Furthermore the linear transformation $\boldsymbol{\xi} = \mathbf{R}_{ij}\boldsymbol{\eta}$ preserves the normal approximation.

Joint Hypothesis Testing

Besides or in addition to pairwise testing, as discussed in the previous section, it may also be of interest to jointly test for the equality of a subset, say, $E_0 \subseteq E$, of partial correlations across groups. This joint hypothesis test can be formulated as

$$H_{0,E_0} : \mathbf{R}_{E_0}\boldsymbol{\rho} = \mathbf{0} \text{ versus } H_{1,E_0} : \mathbf{R}_{E_0}\boldsymbol{\rho} \neq \mathbf{0},$$

where $\mathbf{R}_{E_0}$ denotes a matrix containing the coefficients of the contrasts of interest. For example, in the case of a network with three variables, a researcher could ask whether the edges have equal strength between Variables 1 and 2, and 1 and 3 across groups, the system of equalities under the null hypothesis can be formulated as

$$\mathbf{R}_{E_0}\boldsymbol{\rho} = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \rho_{12,1} \\ \rho_{13,1} \\ \rho_{23,1} \\ \rho_{12,2} \\ \rho_{13,2} \\ \rho_{23,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{24}$$

To quantify the evidence between the null and the alternative hypothesis for the joint test, the same steps can be applied as for the pairwise test where $\mathbf{R}_{E_0}$ replaces $\mathbf{R}_{ij}$ in Step 3. Note that this formulation extends beyond testing two partial correlations. It also applies to testing entire networks (i.e., all edges are the same), or to specific aspects such as invariant edges for a specific node. The latter allows for asking specific questions about network similarity, even when the entire network structure is determined to be different. That is, perhaps there are a priori expectations for relations between specific variables in the network. We demonstrate this approach below (see Application section).

Numerical Performance

The following simulations address two primary aims. The first examined posterior model probabilities with respect to different values for the hyperparameter δ, in addition to how this was influenced by the number of groups tested simultaneously. Although we focus on pairwise hypothesis testing (see Pairwise Hypothesis Testing section), varying the number of groups allows for determining the extent to which the number of hypotheses

under consideration influences the posterior probabilities. The second simulation focuses on error rates and power for detecting edge differences. We do not compare to the NCT method (although edge tests are possible), and instead perform significance testing on the Fisher transformed edge differences estimated with maximum likelihood. This decision was made because it has an analytic solution, which avoids resampling and provides a valuable baseline for comparison.[1] The following used a Bayes factor of 3 as the evidentiary threshold (Kass & Raftery, 1995).

Hyperparameter selection. We used the same partial correlation matrix as in the Null Distribution section (Figure A1). We again focus on the strongest edge in the network ($\rho_{1,3} = 0.46$), which for each simulation trial, was reduced for only one group. This reduction ranged between 0% (i.e., all groups are the same) to 100% (i.e., a difference of 0.46). In other words, for Group 1 and a 75% reduction, data were generated with $\rho_{1,3,g_1} = 0.46 \cdot 0.25$ whereas the generating matrix for the remaining groups was left unaltered ($\rho_{i,j,g\neq 1} = 0.46$). For this simulation we assumed equal sample sizes $n \in \{100 \text{ and } 400\}$, three values for the hyperparamter $\delta \in \{10, 20, \text{ and } 40\}$, which corresponds to prior standard deviations of approximately $s_p \in \{0.30, 0.22, \text{ and } 0.16\}$, and three numbers of groups $G \in \{2, 3, \text{ and } 4\}$. The posterior probabilities in favor of the unrestricted model, that is all groups have the same $\rho_{1,3}$ versus the alternative hypothesis (H_u), were averaged over 100 simulation trials.

These results are presented in Figure 3 (panel A). The *y*-axis denotes the unconstrained model posterior probability for $\rho_{1,3}$. For the *x*-axis a 0% reduction corresponds to the null hypothesis, in that all groups were equal, whereas any amount of reduction resulted in the alternative model being *true* (in this case Group 1 was different). Here the influence of δ can be seen, in particular when the null hypothesis was true, for example the smallest value $\delta = 10$ ($s_p = 0.30$) resulted in the most support for H_0 (i.e., the probability for H_u was the lowest). Further, this difference between hyperparamter values became increasingly pronounced with more hypotheses under consideration. For example, again in reference to the 0% reduction condition, the probability in favor of H_u steadily decreased for $\delta = 10$ as the number of groups increased. On the other hand, for the largest value $\delta = 40$ ($s_p = 0.16$), the average probability was around 0.50 which indicates that it is difficult to gain evidence for the null hypothesis for these sample sizes. A similar pattern was observed when H_u was true, in that largest probabilities were observed for $\delta = 40$. Further context for these results, in reference to error rates and power, is provided below.

Pairwise error rates. In this section we investigate error rates and power for the proposed method. We used the same partial correlation matrix (Figure A1), but this time set values less than 0.10 to zero. This cutoff was chosen to ensure there was adequate power to detect the majority of edge differences in the respective networks. This then served as the covariance structure for Group 1, whereas for the remaining groups it was an identity matrix. Thus, all partial correlations were zero, which created pairwise differences with Group 1. As performance measures, we looked at specificity (SPC) and sensitivity (SN), i.e.,

$$\begin{aligned} \text{SPC} &= \frac{\#\text{true\ negatives}}{\#\text{true\ negatives} + \#\text{false\ positives}}, \\ \text{SN} &= \frac{\#\text{true\ positives}}{\#\text{true\ positives} + \#\text{false\ negatives}}. \end{aligned} \tag{25}$$

The former can be understood in relation to the Type I error rate which is 1 − SPC, and the latter is "power" (1 − SN = the type II error rate). The simulation conditions paralleled the previous section, in that we assumed three values for the hyperparamter $\delta \in \{10, 20, \text{ and } 40\}$ and also three numbers of groups $G \in \{2, 3, \text{ and } 4\}$. We looked at the following sample sizes $n \in \{100, 250, 500 \text{ and } 1{,}000\}$. We could not find any frequentist implementations for jointly testing several correlations. As such the maximum likelihood based method is only included for the two-group condition ($\alpha = .01$). The scores were averaged over 100 simulation trials.

These results are presented in Figure 3 (panel B and C). The performance scores for detecting nonzero edges are displayed in panel B, whereas panel C included the results for detecting zero edges. The latter was accomplished by switching the labels (i.e., 0's changed to 1's) and then computing the scores with (25). Also note that frequentist hypothesis testing (denoted MLE), with $\alpha = .01$, is only included in panel B and for the "two-groups" conditions. All hyperparameter values were competitive with the MLE that, as expected, was calibrated to 99% SPC ($1 - \alpha$). However, the largest value ($\delta = 40$; $s_p = 0.16$) also had the lowest specificity for the smallest sample size and this became pronounced with more groups. Note that the error rate steadily decreased with larger sample sizes, such that all methods performed similarly with larger sample size. On the other hand, when also considering sensitivity ("power"), the MLE was more conservative for the smallest sample sizes while the Bayesian methods were not only able to detect more effects but also had a comparable score for SPC (excluding $\delta = 40$). Finally, for all prior distributions, the Bayes factor showed consistent behavior in that the errors steadily reduced to zero as $n \to \infty$, in addition to increasing scores for SN.

The results for detecting the (true) null hypothesis are provided in panel C. These are particularly important, because they highlight the previously described asymmetry that can arise with too informative or too diffuse prior distributions (Gu, Hoijtink, & Mulder, 2016). For example, with $\delta = 10$ (the least informative prior), SPC was strikingly low for the smallest sample sizes. In other words, the false alarm rate for incorrectly supporting the null hypothesis exceeded 0.50 ($n = 100$). On the other hand, the other hyperparameter values had much higher specificity that improved with the larger sample sizes. Together, when considering sensitivity for detecting nonzero edges, these simulations point toward possible default values for δ. That is, with the explicit goal of balancing the errors for both H_u and H_0, hyperparamter values between 20 and 40 should be used for more than two groups in particular.

Application

We now apply our methods to posttraumatic stress disorder symptoms that were measured in four groups ($N_{g_1} = 526$, $N_{g_2} = 365$, $N_{g_3} = 926$, and $NN_{g_4} = 956$). The symptoms and corresponding node numbers are provided in Table 2. Detailed information about the samples is provided in Fried et al. (2018). The partial correlation matrices are displayed in Figure 4 (panel A). For aesthetic purposes edges smaller than 0.05 were set to zero. Importantly, because the presented methods require the posterior

[1] This was accomplished by computing the difference and the corresponding standard error.

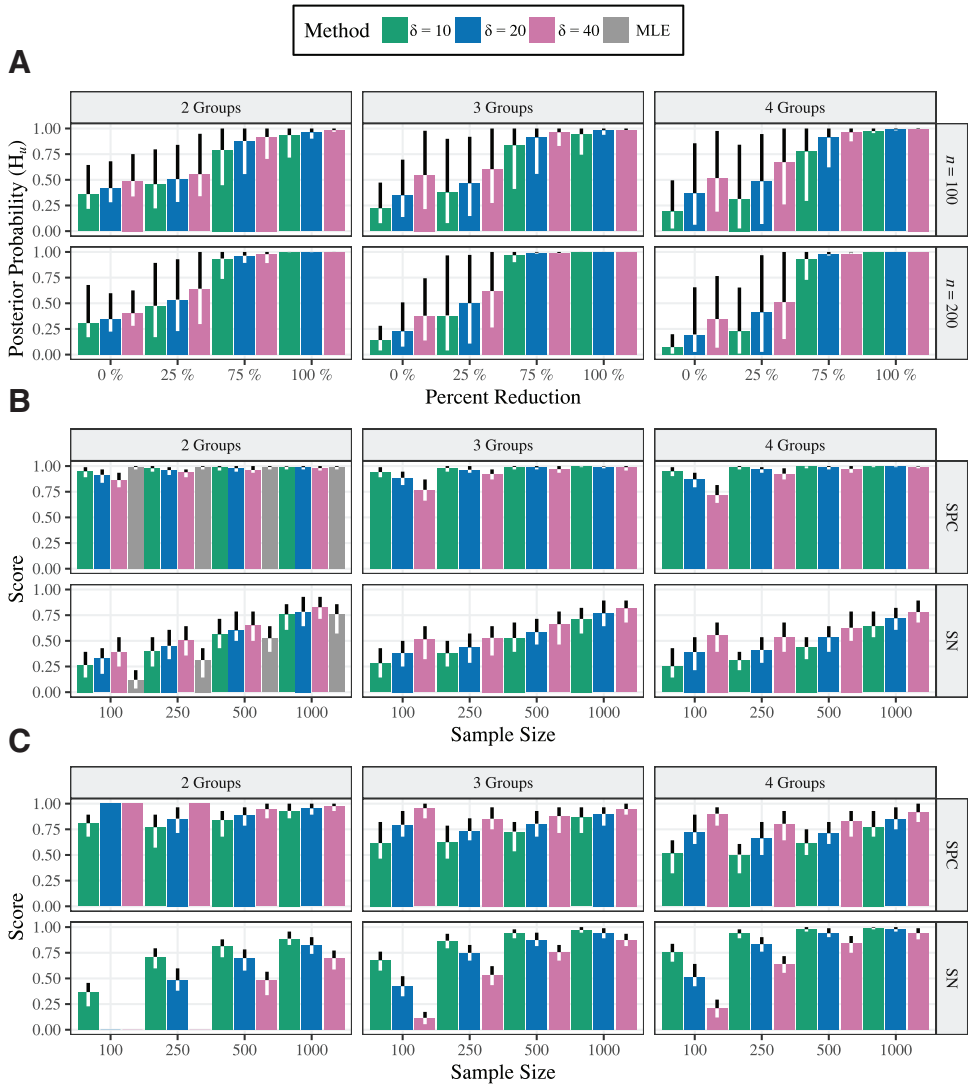

Figure 3. (A) Posterior probabilities for the unconstrained model. Percent reduction is the decrease applied to the maximum edge ($\rho_{1,3,g1} \approx 0.46$) for group number 1. The remaining groups had identical edges for $\rho_{1,3}$. (B) Performance scores for detecting nonzero effects. The MLE corresponds to using confidence intervals with $\alpha = .01$ (only included for the "two-groups" panel). (C) Performance scores for detecting zeros (i.e., the null hypothesis.) SPC = specificity; SN = sensitivity; MLE = maximum likelihood estimate. Error bars are 90% highest density intervals. See the online article for the color version of this figure.

distributions (nothing is set to zero), we emphasize these plots are to visualize the respective edges and not to infer the underlying conditional dependence structures. Further note that we only had access to the correlation matrices, but it is possible to generate data with an empirical (in contrast to population) covariance structure. The following examples are for demonstrative purposes, wherein the intent it primarily to highlight the information provided by the proposed methods.

Posterior Predictive Distribution

We first tested $\mathcal{M}_0$ (5) with the predictive method (posterior predictive distribution). The posterior assuming group equality was computed with all four groups—for example,

$$p\left(\Theta \mid \mathbf{Y}_{g_1}^{obs}, \mathbf{Y}_{g_2}^{obs}, \mathbf{Y}_{g_3}^{obs}, \mathbf{Y}_{g_4}^{obs}, \mathcal{M}_0\right). \tag{26}$$

For each of the 10,000 posterior samples, with the prior given in (6), we then performed pairwise comparisons in which the posterior predictive distribution of $\boldsymbol{\Theta}$ was sampled with the respective samples sizes of the groups being compared. The p values were computed with (4).

The results are displayed in Figure 1 (panel B). For aesthetic purposes the results are presented on the logarithmic scale. The densities correspond to the predictive JSD, that is a symmetric version of Kullback-Leibler divergence (12). The black dots are the observed distances between two multivariate normal distributions, where the density greater than the observed value is the

Table 2
Node Descriptions

Node	Symptom
1	Intrusive thoughts
2	Nightmares
3	Flashbacks
4	Physiological/psychological reactivity
5	Avoidance of thoughts
6	Avoidance of situations
7	Amnesia
8	Disinterest in activities
9	Feeling detached
10	Emotional numbing
11	Foreshortened future
12	Sleep problems
13	Irritability
14	Concentration problems
15	Hypervigilance
16	Startle response

posterior predictive p value. Here it was revealed that $\mathcal{M}_0$ would be rejected at any α level, in that a total of zero predictive draws exceeded the observed distance. In other words, the error for all groups was much greater than that expected under the null model of group equality. These results also parallel the simulation results, in particular the example plot (Figure A1), where the largest groups size had the least amount predictive divergence. Of note the NCT method based on the maximum difference came to a similar conclusion (see: Fried et al., 2018). However, it is important to consider the question asked by each approach. The predictive approach explicitly answers the question of whether two covariance structures, and inversely two precision matrices, were generated from different multivariate normal distributions. This is the necessary assumption behind partial correlations corresponding to conditionally (in)dependent effects (Baba et al., 2004). In the discussion we describe extensions to this approach, for example that essentially any loss function can serve as the discrepancy measure.

We now discuss the results for the nodewise testing approach (see Network Predictive Check section). The node names are provided in Table 3. Furthermore, to make clear what is being tested we have plotted one of the nodes in Figure 4 (panel B). We did not correct the p values (although this would be possible), as our primary focus is to demonstrate the proposed method and the information provided therein. We return to this in the Discussion. However, as a point of reference, $\mathbf{Y}^{rep}_{0.95}$ can be understood as the critical value that corresponds to $\alpha = .05$. It appears that specific groups were different from one another, for example Group 3 and 4, whereas Groups 1 and 2 did not have many small p values. Of course, this could be related to power in that the former also had the largest sample sizes. Interestingly, the only node in which the p value was never smaller than 0.05 was for irritability (i.e., Node 13).

Bayesian Model Comparison

The predictive approach shares some similarities with classical measurement invariance testing, in that failing to reject the null hypothesis does not provide evidence for the null hypothesis. Further, because nothing is fixed (e.g., factor loadings) it also does not provide insight into where the difference is. The following allows for answering more detailed questions about potential differences as well as similarities between network structures. Because $\mathcal{M}_0$ was rejected for all pairwise contrasts, we do not test the

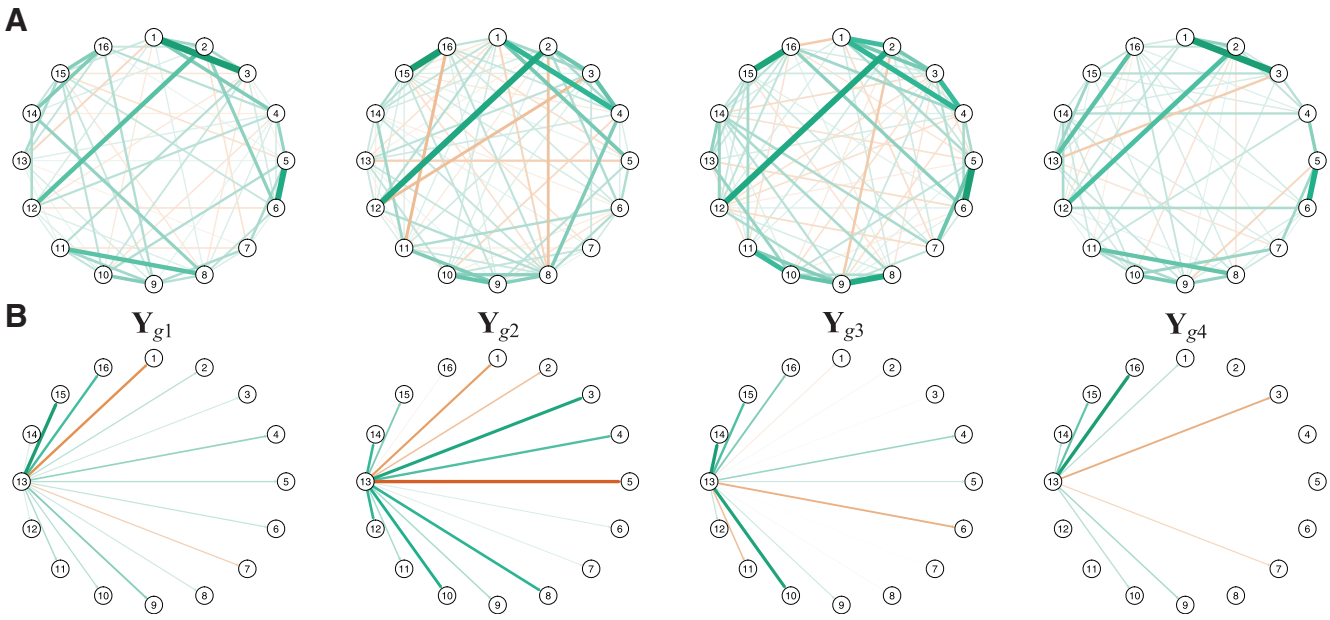

Figure 4. (A) Partial correlation matrices for each group. Values less than 0.05 were set to zero. (B) Example node ("irritability") in the network. Each node was tested with the nodewise predictive method. The null hypothesis of equality was not rejected for this node (see Table 3). See the online article for the color version of this figure.

Table 3
Nodewise Predictive Test

	Y_{g_1} vs. Y_{g_2}			Y_{g_1} vs. Y_{g_3}			Y_{g_1} vs. Y_{g_4}		
Node	$\mathbf{Y}^{rep}_{0.95}$	Y^{obs}	p-value	$\mathbf{Y}^{rep}_{0.95}$	Y^{obs}	p-value	$\mathbf{Y}^{rep}_{0.95}$	Y^{obs}	p-value
1	−5.32	−4.91	0.02	−5.67	−7.01	0.30	−5.75	−3.05	0
2	−4.92	−4.38	0.01	−5.38	−7.65	0.53	−5.39	−2.56	0
3	−5.12	−5.42	0.09	−5.53	−7.56	0.48	−5.56	−3.37	0
4	−5.05	−7.35	0.54	−5.48	−5.72	0.08	−5.50	−2.1	0
5	−4.47	−2.26	0	−4.85	−13.55	0.98	−4.90	−1.81	0
6	−4.66	−2.35	0	−5.08	−9.37	0.83	−5.09	−1.97	0
7	−3.14	−9.8	0.95	−3.49	−1.36	0	−3.49	−6.21	0.62
8	−4.49	−9.45	0.87	−4.93	−13.98	0.98	−4.91	−2.43	0
9	−4.78	−8.8	0.79	−5.18	−3.82	0	−5.21	−4.01	0.04
10	−4.00	−7.37	0.70	−4.38	−4.47	0.06	−4.40	−2.6	0
11	−4.39	−3.72	0.01	−4.85	−8.39	0.75	−4.82	−3.04	0
12	−4.43	−4.01	0.02	−4.81	−8.89	0.80	−4.86	−2.84	0
13	−3.74	−4.82	0.23	−4.21	−6.16	0.46	−4.19	−4.83	0.15
14	−4.38	−4.32	0.04	−4.85	−5.92	0.26	−4.80	−3.19	0
15	−4.41	−5.9	0.34	−4.81	−3.62	0	−4.79	−3.15	0.01
16	−4.60	−5.71	0.25	−5.10	−8.26	0.71	−5.02	−2.51	0
	Y_{g_2} vs. Y_{g_3}			Y_{g_2} vs. Y_{g_4}			Y_{g_3} vs. Y_{g_4}		
Node	$\mathbf{Y}^{rep}_{0.95}$	Y^{obs}	p-value	$\mathbf{Y}^{rep}_{0.95}$	Y^{obs}	p-value	$\mathbf{Y}^{rep}_{0.95}$	Y^{obs}	p-value
1	−5.49	−5.78	0.09	−5.52	−4.06	0	−6.03	−3.35	0
2	−5.14	−4.82	0.02	−5.15	−3.62	0	−5.69	−2.73	0
3	−5.28	−6.27	0.23	−5.29	−4.27	0	−5.86	−3.64	0
4	−5.27	−6.9	0.40	−5.24	−1.95	0	−5.86	−1.77	0
5	−4.61	−2.27	0	−4.68	−5.14	0.12	−5.22	−1.82	0
6	−4.80	−2.29	0	−4.83	−5.55	0.18	−5.42	−1.91	0
7	−3.28	−1.33	0	−3.19	−6.57	0.72	−3.87	−1.17	0
8	−4.66	−9.67	0.88	−4.68	−2.49	0	−5.26	−2.43	0
9	−4.92	−3.99	0	−4.93	−3.84	0	−5.57	−2.51	0
10	−4.19	−4.05	0.04	−4.21	−2.8	0	−4.78	−1.91	0
11	−4.52	−3.93	0.01	−4.57	−5.57	0.23	−5.11	−3.19	0
12	−4.54	−4.2	0.02	−4.60	−4.5	0.04	−5.16	−2.94	0
13	−3.88	−6.25	0.54	−3.93	−15.44	1	−4.50	−6.27	0.42
14	−4.49	−3.58	0	−4.58	−4.90	0.10	−5.15	−2.73	0
15	−4.56	−4.41	0.03	−4.59	−2.69	0	−5.19	−1.95	0
16	−4.74	−6.36	0.38	−4.78	−2.98	0	−5.41	−2.63	0

entire network structure for equality (although this is possible). Instead, again for demonstrative purposes, we focus on individual edges in the networks.

We begin by testing the individual edges for all groups—for example,

$$H_{0,ij}: \rho_{ij,1} = \ldots = \rho_{ij,G} \text{ vs. } H_{1,ij} : \text{ “not } H_{0,ij}\text{”}.$$

The multivariate normal density is then evaluated after applying a linear transformation, which for the posterior mean vector $\boldsymbol{\mu}_{\xi,N}$, follows

$$\mathbf{R}_{ij}\boldsymbol{\rho} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \rho_{ij,g_1} \\ \rho_{ij,g_2} \\ \rho_{ij,g_3} \\ \rho_{ij,g_4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{27}$$

The Bayes factor for each edge, assuming the same transformation has also been applied to the prior distributions, is then given by

$$B_{01,ij} \approx \frac{N(\mathbf{0};\, \boldsymbol{\mu}_{\xi,N}, \Psi_{\xi,N})}{N(\mathbf{0};\, \mathbf{0}, \Psi_{\xi,0})}. \tag{28}$$

In this case the groups are assumed to be independent. For each group we sampled 50,000 draws from the posterior and prior distributions with $\delta = 20$ (i.e., $s_p = 0.22$), and then computed the Bayes factor in (28). We assumed equal prior probabilities for each hypothesis, which is the customary approach for Bayesian hypothesis testing. The results are presented in Figure 5 (panel C), where the Bayes factors are on the logarithmic scale. There was evidence for the null hypothesis of group equality in 52% of the edges. On the other hand, for 30% of the edges there was evidence for the unrestricted model (i.e., a difference). Importantly, because the Bayes factor provides relative evidence, we emphasize this tells us there is more support for “not $H_{0,ij}$” but this is not absolute (i.e., it is restricted to the models under consideration). For the remaining edges the Bayes factors did not exceed the threshold of three. Interestingly, for each node in the network, there were at least two edges for which there was evidence for a difference in strength. Because of this finding, in combination with the posterior predictive results, we decided against investigating further hypotheses. However, note that this general approach applies to essentially any hypothesis one can formulate. We further discuss this in the Discussion.

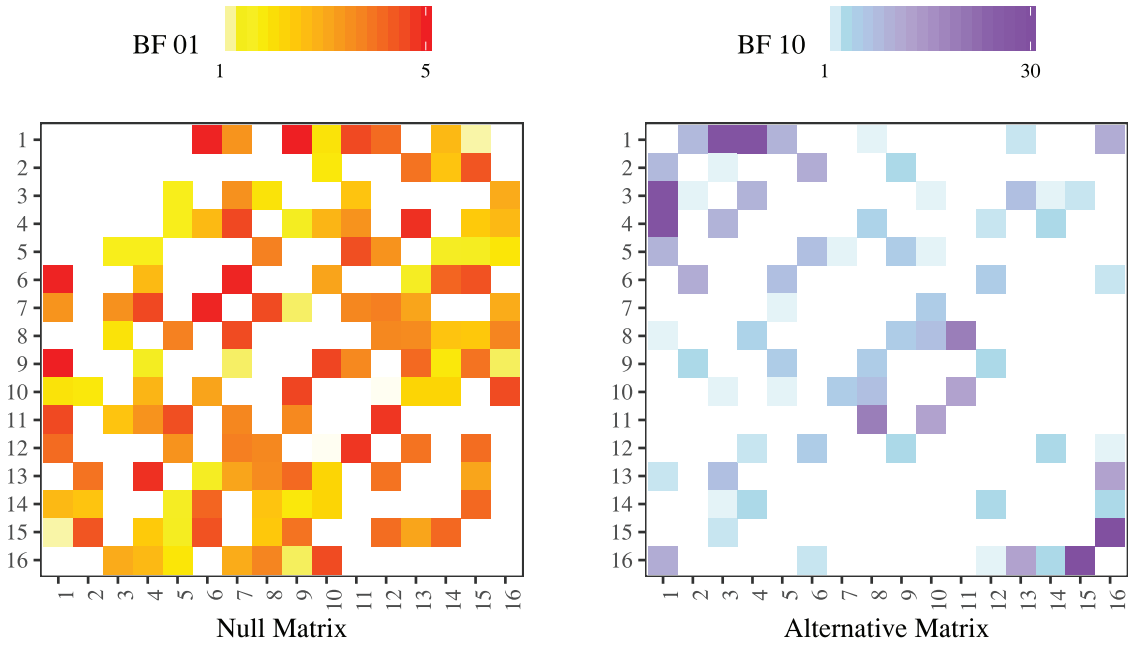

Figure 5. The Bayes factor (BF; on the logarithmic scale) for each individual edge. The null model assumed that each edge was the same in each group. The left plot includes edges for which there was evidence for group equality, whereas in the right plot there was evidence for the alternative hypothesis ("Not H_0"). The empty tiles correspond to a Bayes factor that was less than 3 ($\log(3) \approx 1.10$). See the online article for the color version of this figure.

Discussion

This work introduced two novel methods for comparing any number of Gaussian graphical models. The first is based on the posterior predictive distribution. which as we demonstrated, provides a powerful test against the null hypothesis of group equality. This test is not limited to the overall network structure, but also applies to individual nodes in the network. This allows one to focus on particular variables, for example in the context of psychopathology, examining differences in particular symptoms across networks could be of interest. The second approach uses Bayesian model selection to compare competing theoretical models as they relate to potential differences, or equality, between networks. Alternative hyperparameters for the matrix-F prior were characterized, wherein a range of values emerged as reasonable defaults that can balance both Type I and II errors for the null relative to alternative hypothesis. We applied the methods to posttraumatic stress disorder symptoms measured in four groups. This served to highlight the information provided by the respective methods, in addition to demonstrating another major contribution of this work—the methods apply to any number of groups.

We emphasize that these novel contributions are not restricted to the social-behavioral sciences, but extend to the general Gaussian graphical model literature. Indeed, only recently was there a proposal in the statistics literature to detect differences between precision matrices estimated with Bayesian methods (Bashir, Carvalho, Hahn, & Jones, 2018). However, because this method focused on individual off-diagonal elements of Θ, we decided against contacting the authors for their MATLAB implementation which would then need to be converted to R for general use in psychology. When focusing on specific edges in low-dimensional settings ($p << n$), a valuable comparison in our view is classical hypothesis testing because it will be calibrated to the desired α level (as seen in Figure 3). Their method also used the graphical lasso procedure which has recently been shown to have poor performance in settings common to the network literature in psychology (Williams & Rast, 2018). Furthermore, as we demonstrated, our methods are much more general and not restricted to detecting pairwise edge differences between two groups. They can accurately detect differences between entire precision matrices or specific nodes, as well as flexible Bayesian hypothesis testing that allows for gaining evidence for equality of network structures. These are all novel contributions. Finally, our methods are implemented in the R package BGGM (Williams & Mulder, 2019b).

This work includes two philosophically distinct approaches for statistical inference. The decision to present both methods together is addressed here. In our view the two tests answer different research questions and therefore they complement each other. First the proposed posterior predictive check tests whether there is "enough evidence" in the data to reject the null model of equal network structures across groups. In the case of misfit the challenge is how to extend the null model to better fit the observed data. Second the Bayes factor test can be used to quantify the "relative evidence" in the data for the hypothesis of equal edge strength against an alternative hypothesis that assumes unequal edge strength. The predictive approach has some parallels to classical significance testing (although the predictive distribution is inherently Bayesian), whereas Bayesian model selection is often presented in opposition to such ideas (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). We believe that falsifying an assumed model does have scientific value (Gelman & Shalizi, 2013). Furthermore, there is interesting work that describes the interplay between inference based on estimation and the Bayes

factor (Rouder, Haaf, & Vandekerckhove, 2018). That said, there are two primary reasons we decided to present both approaches. First, because network modeling is relatively new in psychology (Epskamp et al., 2018), there are limited statistical tools available to applied researchers (e.g., compared with SEM). For example, in the case of one network, only recently was an approach for confirmatory hypothesis testing described (Williams & Mulder, 2019a). As such, this work fills a large gap in literature that we viewed as more important than adhering to a particular statistical philosophy.

Second, as we articulated in this work, each approach has different inferential goals. In applied setting this can be advantageous depending on the research question. For example, to investigate misfit from an assumed model, the predictive method provides a powerful test for this purpose. On the other hand, to fully realize the benefits of Bayesian hypothesis testing the hypotheses should be derived from theory (i.e., scientific expectations; Mulder & Olsson-Collentine, 2018). It is unlikely that a theory makes hundreds of predictions, but is rather focused on a subset of edges in the respective networks (see Joint Hypothesis Testing section). In addition to evaluating individual edge differences (as well as invariances) between any number of groups (Figure 3; panel C), we encourage applied researchers to test specific hypotheses in network models. We emphasize that the inferential goal should be decided a priori and the respective hypotheses preregistered. We refer to Faelens, Hoorelbeke, Fried, De Raedt, and Koster (2019) that includes the first preregistered network analysis.

Note that we did not discuss the substantive implications of the applied examples. Furthermore, we will not make specific claims about network replicability based on these data. Nonetheless, in the more general sense, the results to raise some important questions that should be addressed going forward. That is, if researchers genuinely believe that the relations constitute a psychological network, then these four networks are indeed much different than one another (Figure 1; panel C). However, to retain $\mathcal{M}_0$, this would quite literally require drawing two samples from the same multivariate normal distribution. While it is customary to test whether the *true* covariance structure has been fitted (e.g., χ^2 in SEM), this hypothesis is typically rejected at some point. On the other hand, perhaps we do not actually fit *true* models and thus, in a model with hundreds of effects, it is expected that $\mathcal{M}_0$ will be rejected. This is important to consider, going forward, because then the focus should shift from considering "networks" (as a whole) to a subset of the most important partial correlations. For example, as seen in Figure 3 (panel C), there was evidence for group equality for several edges. The methods presented in the work thus allow for testing an ambitious hypothesis (i.e., $\mathcal{M}_0$), in addition to more specific hypotheses about particular nodes (see Table 3), individual edges (Figure 3; panel C), or a subset of edges (see Joint Hypothesis Testing section).

Future Directions

There are Bayesian methods that can jointly estimate Gaussian graphical models (Lin, Wang, Yang, & Zhao, 2015; Peterson, Stingo, & Vannucci, 2015), where information is shared across networks to improve accuracy. This has been shown to lower the false positive and negative rate compared to estimating the networks independently from one another. This is similar to the joint graphical lasso (JGL) that is commonly used in psychology. Indeed, it was used to jointly estimate the conditional dependence structures the four data sets used in this work (Fried et al., 2018). However, we would caution applied researchers from assuming methods like the JGL accurately estimate psychological networks (e.g., compared to independently with Bayesian or maximum likelihood estimation; Williams & Rast, 2018). The simulation conditions in Danaher, Wang, and Witten (2014), where the JGL was introduced, were not representative of the psychological network literature (e.g., $p = 1000$ and $n = 100$). As such, it is not clear whether the reported advantages extend to more common situations in the social-behavioral sciences ($p < n$). Nonetheless, it would be interesting to extend the present methods to jointly estimate the conditional dependence structures of (potentially) any number of networks. Here it could be determined if there are indeed advantage compared independent estimation that was shown to have excellent performance in this work (i.e., Figures 1 and 3).

Additionally, the posterior predictive method is not limited to KL-divergence such that any test statistic could be used as the discrepancy measure. To parallel the NCT package (van Borkulo et al., 2017) it would be possible to obtain the predictive distribution of absolute error between partial correlations matrices. However, we would not limit the possibilities to this current article or what is implemented in the NCT package. For example, a measure that is related to binary classification such as Hamming distance (Norouzi, Fleet, & Salakhutdinov, 2012) or Matthews correlation coefficient which is a measure of association for binary variables (e.g., adjacency matrices, Powers, 2011). However, before employing an alternative measure in practice, its numerical performance should first be evaluated to understand its frequentist properties (Rubin, 1984).

On the other hand, we know more about the properties of Bayesian model selection (Casella, Girón, Martinez, & Moreno, 2009)—that is, the Bayes factor is known to converge on the *true* model with infinite data. As such, the package BGGM includes approaches that extend beyond what is presented in this work. It is possible to test any hypothesis of interest. In the context of an experimental design (control vs. treatment), one possibility is that a subset of edges stayed the same, others increased, while yet others decreased in response to the treatment. This can be tested with the method described in the Joint Hypothesis Testing section. Because our focus was on introducing two novel methods, it was beyond the scope of this work to provide more detailed instruction (although there are examples in the package documentation). Consequently, we plan to write an in-depth tutorial that applies Bayesian model comparison to test specific hypotheses of interest in network models.

Limitations

There are limitations of this work. First, because network models include several edges (typically over 100), determining how best to evaluate numerical performance was not straightforward. The simulation conditions, in this regard, were simplified to focus on key aspects of the proposed methods—for example, demonstrating calibrated error rates under the null hypothesis (see Table 1). However, the predictive distribution and Bayesian hypothesis testing are well established approaches in the Bayesian literature. As such, there is no reason to assume that the known properties of each would not extend to Gaussian graphical models (especially when there is a direct correspondence to multiple regression;

Kwan, 2014; Stephens, 1998). Examining performance, going forward, would be particularly important in the context of model misspecification (e.g., omitted nodes).

Second, we did not consider estimating the conditional (in)dependence structures. We refer interested readers to Williams and Mulder (2019a), where Bayesian methods specifically for determining the edge set in one network are described. These are also implemented in the package BGGM. Moreover, because the focus of this work was explicitly on low-dimensional settings, we considered it a given that the models would be accurately estimated. Relatedly, note that in a Bayesian context there is never a truly sparse solution and thus a decision rule is required for determining the edge set. However, when considering differences between networks, this can be advantageous because no postprocessing is required. The method described in Belilovsky et al. (2016) first used l_1-regularization and then desparsified the estimates after the fact (Van De Geer, Bühlmann, Ritov, & Dezeure, 2014). This removes the zeroes, which then allows for constructing confidence intervals to conduct classical significance tests on the respective differences. Of course, this is entirely unnecessary because confidence intervals can readily be constructed from nonregularized partial correlations (as done in this work, which assumes $p < n$; Williams & Rast, 2018). Similarly, while not included here, it would be straightforward to subtract the posterior distributions for two edges and then check the credible interval for zero. In contrast to using the Bayes factor, this cannot provide evidence for the null hypothesis. In the case of the predictive method, note that imposing zeroes would alter the joint posterior density, thereby resulting in a distorted predictive distribution.

Third it is well-known that the Bayes factor is sensitive to the prior standard deviation of the effect under the alternative. This was also observed in this work through the choice of δ in our parameterization of the matrix-*F* prior distribution. This however is not necessarily a negative property because it forces the researcher to carefully think about the anticipated effect, through δ, if the null model would be false. Although specifying δ may be difficult, especially because the network approach is relatively new in psychological science, we expect that network researchers are able to make sensible choices for the prior standard deviation of the effect under the alternative based on their own prior experience or based on results from published literature. In the case of prior uncertainty, it is recommended to perform a prior sensitivity analysis by computing the Bayes factor based on (realistic) minimal and maximal anticipated effects. This would provide a realistic range of the relative evidence in the data between the hypotheses of interest.

Fourth, although it would be possible to adjust to the posterior predictive *p* values (e.g., controlling false discovery rate; Benjamini & Hochberg, 1995), this will not always be possible. This is due to the fact that the *p* value can be exactly zero, wherein none of the predictive draws exceed the observed distance (see Table 3). This indicates a substantial difference from what the null model predicts but should be considered nonetheless. Alternatively, it is perfectly acceptable to interpret the *p* values as a continuous measure of discrepancy from the assumed model (i.e., of group equality; Greenland, 2017). We prefer this approach in practice, and emphasize the thresholds used in this work (i.e., $\alpha = .05$ and $B_{01} > 3$) were necessarily adopted to evaluate numerical performance.

Lastly, this work focused exclusively on continuous data. It is common in psychology to have ordinal data, for example constructs measured with Likert scales. While it was shown that assuming normality for five-level ordinal had close to nominal error rates in networks, which parallels (Rhemtulla, Brosseau-Liard, & Savalei, 2012), we caution against using these methods for ordinal data with few categories. We plan to extend these methods to allow for comparing polychoric partial correlations between groups.

Conclusion

We introduced two novel methods for comparing Gaussian graphical models. The applied examples demonstrated the utility of the proposed methods. They can be used to test the null hypothesis of network equality, or gain evidence for invariant network structures with the Bayes factor. To ensure the methods can readily be adopted by applied researchers, they are implemented in the R package BGGM.

References

Baba, K., Shibata, R., & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics, 46,* 657–664. http://dx.doi.org/10.1111/j.1467-842X.2004.00360.x

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modelling covariance matrices in terms of standard deviations and correlations with applications to shrinkage. *Statistica Sinica, 10,* 1281–1311.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika, 44,* 533–534. http://dx.doi.org/10.1093/biomet/44.3-4.533

Bashir, A., Carvalho, C. M., Hahn, P. R., & Jones, M. B. (2018). Post-processing posteriors over precision matrices to produce sparse graph estimates. *Bayesian Analysis, 14,* 1–16. http://dx.doi.org/10.1214/18-ba1139

Bayarri, M. J., & Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association, 95,* 1127–1142. http://dx.doi.org/10.1080/01621459.2000.10474309

Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M. T., & Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine, 46,* 3359–3369. http://dx.doi.org/10.1017/S0033291716002300

Belilovsky, E., Varoquaux, G., & Blaschko, M. B. (2016). Testing for differences in Gaussian graphical models: Applications to brain connectivity. Advances in neural information processing systems (pp. 595–603). New York, NY: Curran Associates Inc.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57,* 289–300.

Bernardo, J. M. (2005). Reference analysis. *Handbook of statistics, 25,* 17–90.

Bernardo, J. M., & Smith, A. F. M. (2001). *Bayesian theory*. New York, NY: NY IOP Publishing.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112,* 859–877. http://dx.doi.org/10.1080/01621459.2017.1285773

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

Cramer, A. O., & Borsboom, D. (2015). Problems attract problems: A network perspective on mental disorders. *Emerging Trends in the Social and Behavioral Sciences*. Advance online publication. http://dx.doi.org/10.1002/9781118900772

Dahl, F. A., Gasemyr, J., & Natvig, B. (2007). A robust conflict measure of inconsistencies in Bayesian hierarchical models. *Scandinavian Jour-*

nal of Statistics. Advance online publication. http://dx.doi.org/10.1111/j.1467-9469.2007.00560.x

Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 76,* 373–397. http://dx.doi.org/10.1111/rssb.12033

Dempster, A. (1972). Covariance selection. *Biometrics, 28,* 157–175. http://dx.doi.org/10.2307/2528966

Deng, H., & Wickham, H. (2011). *Density estimation* in R. Retrieved from http://www2.cs.uh.edu/~ceick/7362/T2-4.pdf

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42,* 204–223. http://dx.doi.org/10.1214/aoms/1177693507

Eguchi, S., & Copas, J. (2006). Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis, 97,* 2034–2040. http://dx.doi.org/10.1016/j.jmva.2006.03.007

Epskamp, S. (2016). Regularized Gaussian psychological networks: Brief report on the performance of extended BIC model selection. *arXiv preprint arXiv:1606.05771*.

Epskamp, S., & Fried, E. I. (2016). A tutorial on regularized partial correlation networks. *arXiv*. http://dx.doi.org/10.1103/PhysRevB.69.161303

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82,* 904–927. http://dx.doi.org/10.1007/s11336-017-9557-x

Epskamp, S., Waldorp, L. J., Mottus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research, 53,* 453–480. http://dx.doi.org/10.1080/00273171.2018.1454823

Faelens, L., Hoorelbeke, K., Fried, E., De Raedt, R., & Koster, E. H. (2019). Negative influences of Facebook use through the lens of network analysis. *Computers in Human Behavior, 96,* 13–22. http://dx.doi.org/10.1016/j.chb.2019.02.002

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

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10,* 507–521.

Fisher, R. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2019). Quantifying the reliability and replicability of psychopathology network characteristics. *Multivariate Behavioral Research*. Advance online publication. http://dx.doi.org/10.1080/00273171.2019.1616526

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Dijk, H. M. H.-v., Bockting, C. L. H., . . . Karstoft, K.-I. (2018). Replicability and generalizability of posttraumatic stress disorder (PTSD) networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples. *Clinical Psychological Science, 6,* 335–351. http://dx.doi.org/10.1177/2167702617745092

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society, 1,* 1–14.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics, 7,* 2595–2602. http://dx.doi.org/10.1214/13-EJS854

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, B. D., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton: CRC Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–807.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66,* 8–38. http://dx.doi.org/10.1111/j.2044-8317.2011.02037.x

Goutis, C. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika, 85,* 29–37. http://dx.doi.org/10.1093/biomet/85.1.29

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology, 186,* 639–645. http://dx.doi.org/10.1093/aje/kwx259

Grewal, M. S. (2011). Kalman filtering. In Miodrag Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 705–708). Berlin, Heidelberg: Springer.

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology, 72,* 130–143. http://dx.doi.org/10.1016/j.jmp.2015.09.001

Hjort, N. L., Dahl, F. A., & Steinbakk, G. H. (2006). *Post-processing posterior predictive p values*. New York, NY: Taylor & Francis. http://dx.doi.org/10.1198/016214505000001393

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphical models with R*. http://dx.doi.org/10.1007/978-1-4614-2299-0

James, W., & Stein, C. (1961). Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics (Vol. 1, pp. 361–379).

Jeffreys, H. (1961). *The theory of probability*. Oxford, UK: Oxford University Press.

Jones, P. J., Williams, D. R., & McNally, R. J. (2019). Sampling variability is not nonreplication: A Bayesian Reanalysis of Forbes, Wright, Markon, & Krueger. *PsyArXiv*. http://dx.doi.org/10.31234/OSF.IO/EGWFJ

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica, 59,* 57–69. http://dx.doi.org/10.1111/j.1467-9574.2005.00279.x

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

Kwan, C. C. Y. (2014). A regression-based interpretation of the inverse of the sample covariance matrix. *Spreadsheets in Education, 7,* 4613.

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

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 519–537. http://dx.doi.org/10.1177/0146621608329504

Lin, Z., Wang, T., Yang, C., & Zhao, H. (2015). *On joint estimation of Gaussian graphical models for spatial and temporal data*. Retrieved from http://arxiv.org/abs/1507.01933. http://dx.doi.org/10.1111/biom.12650

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44,* 187–192. http://dx.doi.org/10.1093/biomet/44.1-2.187

Marathe, A., Pan, Z., & Apolloni, A. (2013). Analysis of friendship network and its role in explaining obesity. *ACM Transactions on Intelligent Systems and Technology (TIST), 4,* 3–21.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22,* 1142–1160. http://dx.doi.org/10.1214/aos/1176325622

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology, 72,* 104–115. http://dx.doi.org/10.1016/j.jmp.2014.09.004

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference, 140,* 887–906. http://dx.doi.org/10.1016/j.jspi.2009.09.022

Mulder, J., & Olsson-Collentine, A. (2018). *Simple Bayesian testing of scientific expectations in linear regression models*. Behavior research methods, *51,* 1117–1130.

Mulder, J., & Raúl Pericchi, L. (2018). The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis, 4,* 1–22. http://dx.doi.org/10.1214/17-BA1092

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods and Research, 47,* 637–664. http://dx.doi.org/10.1177/0049124117701488

Nielsen, F. (2010). *A family of statistical symmetric divergences based on Jensen's inequality. arXiv preprint arXiv:1009.4004*. Retrieved from https://arxiv.org/abs/1009.4004

Norouzi, M., Fleet, D. J., & Salakhutdinov, R. R. (2012). Hamming distance metric learning. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 1061–1069). New York, NY: Curran Associates, Inc.

Padmanabhan, N., White, M., Zhou, H. H., & O'Connell, R. (2016). Estimating sparse precision matrices. *Monthly Notices of the Royal Astronomical Society, 460,* 1567–1576. http://dx.doi.org/10.1093/mnras/stw1042

Peterson, C., Stingo, F. C., & Vannucci, M. (2015). Bayesian Inference of multiple Gaussian graphical models. *Journal of the American Statistical Association, 110,* 159–174. http://dx.doi.org/10.1080/01621459.2014.896806

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing, 27,* 711–735. http://dx.doi.org/10.1007/s11222-016-9649-y

Powers, D. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2,* 37–63. http://dx.doi.org/10.9735/2229-3981

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17,* 1–14. http://dx.doi.org/10.1037/a0026804

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. http://dx.doi.org/10.1037/a0029315

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association, 95,* 1143–1156. http://dx.doi.org/10.1080/01621459.2000.10474310

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin and Review, 25,* 102–113. http://dx.doi.org/10.3758/s13423-017-1420-7

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12,* 1151–1172. http://dx.doi.org/10.1214/aos/1176346785

Sinharay, S., & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference, 111,* 209–221. http://dx.doi.org/10.1016/S0378-3758(02)00303-8

Stephens, G. (1998). On the inverse of the covariance matrix in portfolio analysis. *The Journal of Finance, 53,* 1821–1827.

van Borkulo, C. D., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2017). Waldorp. http://dx.doi.org/10.13140/RG.2.2.29455.38569

Van De Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics, 42,* 1166–1202. http://dx.doi.org/10.1214/14-AOS1221

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9,* 486–492. http://dx.doi.org/10.1080/17405629.2012.686740

van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6,* 1–4. http://dx.doi.org/10.3389/fpsyg.2015.01064

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2017). Posterior calibration of posterior predictive p values. *Psychological Methods, 22,* 382–396. http://dx.doi.org/10.1037/met0000142

Verhagen, J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology, 66,* 383–401. http://dx.doi.org/10.1111/j.2044-8317.2012.02059.x

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology, 72,* 171–182. http://dx.doi.org/10.1016/j.jmp.2015.06.005

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. http://dx.doi.org/10.1037/a0022790

Werner, M., Štulhofer, A., Waldorp, L., & Jurin, T. (2018). A network approach to hypersexuality: insights and clinical implications. *Journal of Sexual Medicine, 15,* 410–415. http://dx.doi.org/10.1016/j.jsxm.2018.01.009

Wetzels, R., Grasman, R. P., & Wagenmakers, E. J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio. *Computational Statistics and Data Analysis, 54,* 2094–2102. http://dx.doi.org/10.1016/j.csda.2010.03.016

Williams, D. R. (2018). Bayesian inference for Gaussian graphical models: Structure learning, explanation, and prediction. *PsyArXiv*. Advance online publication. http://dx.doi.org/10.31234/OSF.IO/X8DPR

Williams, D. R., & Mulder, J. (2019a). Bayesian hypothesis testing for gaussian graphical models: Conditional independence and order constraints. *PsyArXiv*. Advance online publication. http://dx.doi.org/10.31234/osf.io/ypxd8

Williams, D. R., & Mulder, J. (2019b). BGGM: A R Package for Bayesian Gaussian graphical models. *PsyArXiv*. Advance online publication. http://dx.doi.org/10.31234/osf.io/3b5hf

Williams, D. R., & Rast, P. (2018). Back to the basics: Rethinking partial correlation network methodology. Advance online publication. http://dx.doi.org/10.31219/osf.io/fndru

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

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Analysis, *13,* 917–1003.

Yule, G. U. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London Series A, 79,* 182–193. http://dx.doi.org/10.1098/rspa.1907.0028

(*Appendices follow*)

Appendix A

Gibbs Sampler

For a precision matrix $\boldsymbol{\Theta}$ having a matrix-$F(\nu, \delta, \mathbf{B})$ prior distribution, the probability density function is given by

$$p^F(\Theta) = \frac{\Gamma_k\left(\frac{\nu+\delta+p-1}{2}\right)}{\Gamma_k\left(\frac{\nu}{2}\right)\Gamma_k\left(\frac{\delta+p-1}{2}\right)|\mathbf{B}|^{\frac{\nu}{2}}}|\Theta|^{\frac{\nu-p-1}{2}}|\mathbf{I}_k+\Theta\mathbf{B}^{-1}|^{-\frac{\nu+\delta+p-1}{2}}. \tag{29}$$

Following Mulder and Pericchi (2018), this implies that the covariance matrix $\boldsymbol{\Sigma}$ follows a matrix-$F(\delta + p - 1, \nu - p + 1, \mathbf{B}^{-1})$, which can be written as a scale mixture of inverse Wishart distributions, i.e.,

$$\begin{aligned}\Sigma \mid \Psi &\sim IW(\nu, \Psi)\\ \Psi &\sim W(\delta + p - 1, \mathbf{B}^{-1}).\end{aligned} \tag{30}$$

Furthermore, the likelihood for n independent observations from the multivariate normal model is given by

$$p(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{n/2}\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}\exp\left\{-\frac{n}{2}(\mu-\bar{\mathbf{y}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{y}})\right\}, \tag{31}$$

where $\mathbf{S}$ is the sums of squares matrix and $\bar{\mathbf{y}}$ is the vector of sample means. Therefore when using a flat prior for the nuisance parameter $\boldsymbol{\mu}$, the posterior (including the parameter expansion with $\boldsymbol{\Psi}$) can be written as

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Psi} \mid \mathbf{Y}) &\propto p^{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi})p^W(\boldsymbol{\Psi})p(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\\ &\propto |\boldsymbol{\Psi}|^{(\nu+\delta-2)/2}|\boldsymbol{\Sigma}|^{-(n+\nu+p+1)/2}\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right\}\\ &\quad\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Psi}\mathbf{B}\right\}\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}\\ &\quad\exp\left\{-\frac{n}{2}(\boldsymbol{\mu}-\bar{\mathbf{y}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{y}})\right\}\\ \Rightarrow p(\boldsymbol{\Sigma}, \boldsymbol{\Psi} \mid \mathbf{Y}) &\propto |\boldsymbol{\Psi}|^{(\nu+\delta-2)/2}|\boldsymbol{\Sigma}|^{-(n+\nu+p)/2}\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right\}\\ &\quad\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Psi}\mathbf{B}\right\}\exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}.\end{aligned}$$

Hence a Gibbs sampler can then be formed using the following conditional posteriors

$$\Sigma \mid \boldsymbol{\Psi}, \mathbf{Y} \sim IW(n + \nu - 1, \mathbf{S} + \boldsymbol{\Psi})$$

$$\boldsymbol{\Psi} \mid \boldsymbol{\Sigma}, \mathbf{Y} \sim W(\nu + \delta + p - 1, (\boldsymbol{\Sigma}^{-1} + \mathbf{B})^{-1}).$$

A posterior sample for the covariance matrix $\boldsymbol{\Sigma}$, and thus of the precision matrix $\boldsymbol{\Theta}$ and the partial correlations, can be obtained by iteratively sampling $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ from their respective conditional posterior distribution.

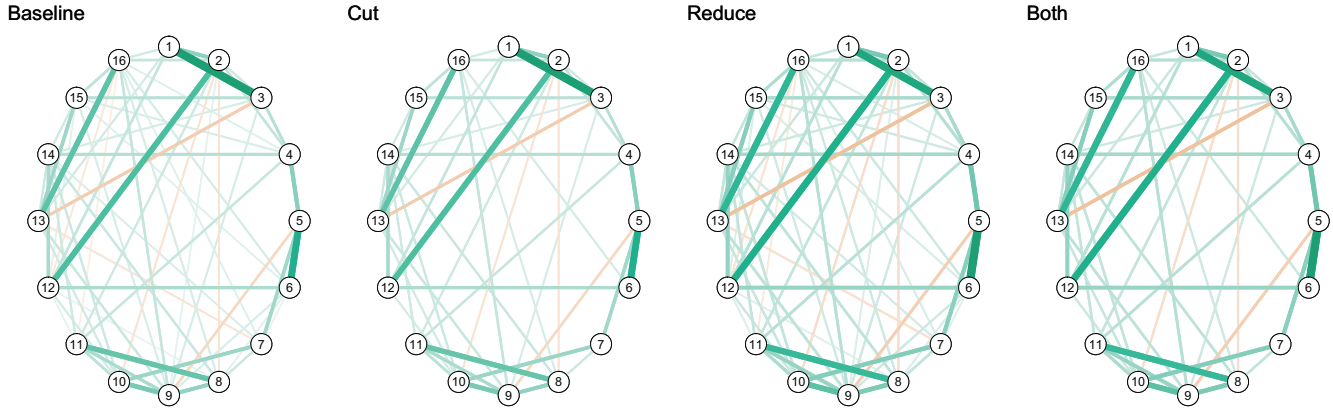

Figure A1. Graphical structures used in the simulation (see Detecting Differences section). Baseline: edges less than 0.05 set to zero. Cut: edges less than 0.075 set to zero. Reduce: largest edge reduced by 25%. Both: edges cut and the largest reduced by 25%. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix B

Timing Results

	Sample Size		
Method	250	500	1,000
Predictive	1.34 (0.01)	1.84 (0.06)	2.73 (0.06)
Permutation	22.40 (0.39)	22.40 (0.25)	22.60 (0.25)

Received February 27, 2019
Revision received October 1, 2019
Accepted November 26, 2019 ■