Bayesian Estimation and Comparison of Idiographic Network Models

Björn S. Siepe[1], Matthias Kloft[1], and Daniel W. Heck[1]

[1]Psychological Methods Lab, Department of Psychology, University of Marburg

Author Note

Björn S. Siepe https://orcid.org/0000-0002-9558-4648

Matthias Kloft https://orcid.org/0000-0003-1845-6957

Daniel W. Heck https://orcid.org/0000-0002-6302-9252

The authors made the following contributions. BS: Conceptualization, Methodology, Formal Analysis, Software, Visualization, Writing - original draft, Writing - review & editing; MK: Methodology, Formal Analysis, Software, Writing - review & editing; DWH: Conceptualization, Methodology, Formal Analysis, Supervision, Writing - review & editing.

Data and R code for all analyses are available at the Open Science Framework: https://osf.io/9byaj/. This study was not preregistered. This work was presented at the bi-annual meeting of the Quantitative Methods Section of the German Psychological Assocation (2023).

This is the second version of this preprint (February 5, 2024). During the revision of this manuscript, Matthias Kloft was added as an author.

We thank Maarten Marsman for helpful discussions on our approach, Raphael Hartmann for helpful comments on a prior version of the manuscript, and Aaron Fisher for making the data used in the empirical example publicly available.

Correspondence concerning this article should be addressed to Björn S. Siepe, Psychological Methods Lab, Department of Psychology, University of Marburg, Gutenbergstraße 18, Marburg, Germany.

E-mail: bjoern.siepe@uni-marburg.de

Abstract

Idiographic network models are estimated on time-series data of a single individual and allow researchers to investigate person-specific associations between multiple variables over time. The most common approach for fitting graphical vector autoregressive (GVAR) models uses LASSO regularization to estimate a contemporaneous and a temporal network. However, estimation of idiographic networks can be unstable in relatively small data sets typical for psychological research. This bears the risk of misinterpreting differences in estimated networks as spurious heterogeneity between individuals. As a remedy, we evaluate the performance of a Bayesian alternative for fitting GVAR models that allows for regularization of parameters while accounting for estimation uncertainty. We also develop a novel test, implemented in the `tsnet` package in R, which assesses whether differences between estimated networks are reliable based on matrix norms. We first compare Bayesian and LASSO approaches across a range of conditions in a simulation study. Overall, LASSO estimation performs well, while a Bayesian GVAR without edge selection may perform better when the true network is dense. In an additional simulation study, the novel test is conservative and shows good false-positive rates. Finally, we apply Bayesian estimation and testing in an empirical example using daily data on clinical symptoms for 40 individuals. We additionally provide functionality to estimate Bayesian GVAR models in Stan within `tsnet`. Overall, Bayesian GVAR modelling facilitates the assessment of estimation uncertainty which is important for studying inter-individual differences of intra-individual dynamics. In doing so, the novel test serves as a safeguard against premature conclusions of heterogeneity.

Keywords: Time series analysis, network analysis, dynamic network, Bayesian estimation, idiographic

Introduction

The idea that symptoms of mental disorders form a network of causally mutually interacting variables has gained popularity as an alternative to classical conceptualizations of psychopathology (Borsboom, 2017). Alongside spurring theoretical debates, the network perspective has inspired many methodological innovations and has been applied in several areas of psychology (Borsboom et al., 2021). Early applications of network modelling in psychology mostly focused on cross-sectional data (e.g., Epskamp, Borsboom, and Fried, 2018). Theoretical and empirical work has since clearly shown the need for a more person-specific perspective to study psychological constructs (Fisher et al., 2018; Hamaker, 2012; Molenaar, 2004). Therefore, the use of idiographic network models for time-series data of a single individual has become a prominent area of research (Bringmann, 2021) to model *dynamic* associations between variables *within* an individual *over time*.

These idiographic approaches are particularly interesting for clinical research due to their potential to provide a tailored perspective on individual case studies and their alignment with the highly individualized perspective in clinical practice (Piccirillo et al., 2019). Although idiographic approaches focus on the data of a single person for network estimation, researchers are often interested in comparing network structures between individuals to investigate inter-individual differences (e.g., Levinson et al., 2022) or within individuals over time to study within-person stability (e.g., Beck & Jackson, 2020). While there is a great theoretical and intuitive appeal to such lines of research, simulation studies have shown that in typical psychological data with relatively few time points, it is difficult to recover the true parameters of a network (Mansueto et al., 2023) and to obtain proper uncertainty measures for estimated parameters. This is due to the considerable sampling variability in small data sets which makes the interpretation of network comparisons very difficult. As a consequence, researchers may erroneously interpret random fluctuations as true differences between networks (Hoekstra et al., 2022). Before turning to solutions to this problem, we first explain how idiographic networks are typically defined and estimated.

The foundation of most idiographic network models is the lag-1 vector autoregressive (VAR) model, in which each variable is regressed on itself and on all other variables at the previous point in time to obtain directed estimates of the *temporal* association between psychological constructs. In typical psychological networks, *nodes* represent variables and *edges* statistical associations between nodes.[1] However, many psychological effects presumably occur faster than the chosen sampling frequency (e.g., days or several hours). Therefore, the *GVAR* approach by Epskamp, van Borkulo, et al. (2018) also models the innovation structure at each time point which resembles the residuals not captured by the (vector-)autoregressive structure (Epskamp, 2020). The innovations are used to obtain a *contemporaneous* network, which captures undirected effects between variables that occur faster than the lag interval of the temporal network, assuming that the model is correctly specified. To estimate the combined model, the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1996) is used in the `graphicalVAR` package (Epskamp & Asena, 2021) to estimate both network structures jointly and to shrink small coefficients to zero. Hereafter, we use the term GVAR to refer to idiographic models consisting of a temporal and a contemporaneous network. Moreover, we refer to LASSO GVAR as its implementation in the `graphicalVAR`-package.[2]

By implementing an automatic edge selection, LASSO regularization can improve interpretability and increase the specificity of estimated networks. However, it also has some drawbacks which have been discussed extensively in the cross-sectional network literature (e.g., Williams et al., 2019). Similar issues apply to network models for longitudinal data. In psychology, the characteristics of the data may differ significantly from other fields where LASSO is commonly used to estimate networks with considerably more nodes (Williams et al., 2019). This raises questions about the extent to which the advantages of LASSO are

[1] There are continuous-time alternatives to estimating network models, but we focus on discrete-time models here and refer to the limitations section for shortcomings of this approach.

[2] Note that there are also other approaches to obtain personalized networks, such as structural VAR modeling (Epskamp, van Borkulo, et al., 2018; Ye et al., 2021), but we focus on the GVAR approach here due to its popularity and its reliance on regularization.

applicable in psychological research. In the longitudinal setting, previous simulation studies have shown that an acceptable performance of idiographic networks with typical psychological data sets is often only possible with a small number of nodes (around six nodes; Mansueto et al., 2023). The benefits of using LASSO regularization for such models with relatively few parameters are unclear. Moreover, the use of LASSO prohibits a simple construction of confidence intervals, since regularized estimates have a point mass at zero in their sampling distribution (Williams et al., 2019). An assessment of uncertainty is thus usually limited to bootstrapping.

Besides the difficulty in quantifying estimation uncertainty, the large influence of sampling variability in small sample sizes may hinder the proper interpretation of estimated networks. In a recent simulation study, Hoekstra et al. (2022) showed that, due to sampling variability, idiographic networks of different individuals may often appear to be more heterogeneous than they actually are. Graphical representations of LASSO-based networks may give a false impression of qualitative differences merely because different edges are set to zero. This may in turn lead researchers to draw potentially incorrect conclusions about the amount of heterogeneity in their sample. Similar issues concerning the uncertainty of estimated networks have led Marsman and Rhemtulla (2022, p. 4) to conclude that *'the robustness of network results now firmly ranks as one of the field's top priorities.'*

Motivated by this goal, we describe a Bayesian GVAR approach, a Bayesian alternative to LASSO which accounts for the uncertainty in estimated idiographic networks. To the best of our knowledge, Bayesian approaches for estimating GVAR models have not been evaluated in the psychological literature yet. In doing so, we first aim to propose a solution to the problem that estimation uncertainty is generally difficult to account for in these models. Second, by using an approach that imposes some regularization but does not aim at setting edges to zero, we offer new tools for researchers who do not wish to invoke sparsity assumptions. Third, we use the information about estimation uncertainty in the posterior samples to develop a novel test for evaluating and testing differences in network models. As a remedy for the issues plaguing network comparisons (Hoekstra et al., 2022), we use samples of the posterior distributions to test whether differences between two estimated networks indeed reflect genuine differences between individuals and not only mere sampling variability. Developing a test of network differences is important to provide researchers with a safeguard against spurious heterogeneity. By harnessing the strengths of Bayesian estimation, we hope to make idiographic network estimation more robust and facilitate the assessment of uncertainty in GVAR modeling.

In the following, we briefly highlight advantages of Bayesian inference relevant to network analysis. However, we do not aim to provide a full introduction to Bayesian modelling and refer the reader to suitable introductions (e.g., van de Schoot et al., 2021, van de Schoot et al., 2017). Bayesian modeling allows us to make probabilistic statements about the parameters in a model by combining prior expectations and information in the observed data (van de Schoot et al., 2021). Uncertainty about the parameters is explicitly accounted for in all steps of Bayesian modelling and is described by the posterior distribution. The posterior distribution also allows us to quantify uncertainty about derived quantities of interest, such as network centrality, or about differences between edges (Williams, 2021). Accounting for uncertainty is particularly important when applying complex models to noisy data for clinical use. In addition, prior distributions can be used to induce shrinkage for estimates or set them to zero, similar to popular regularization approaches such as LASSO or Ridge regression, while still yielding measures of uncertainty for the estimated parameters (van Erp et al., 2019). Bayesian estimation also facilitates the inclusion of expert opinions or results of previous studies into the estimation of networks via the prior distribution. For example, the *PREMISE*-framework by Burger et al. (2022) elicits therapists' opinions about the direction and strength of the associations between their clients' symptoms. This can then be used as a prior distribution for networks estimated from time series data. In summary, Bayesian methods offer a number of benefits that have not yet been leveraged in GVAR models.

Our paper is structured as follows. In Part 1, we illustrate the Bayesian estimation of idiographic networks. In a simulation study, we evaluate the performance of Bayesian inference in different settings and compare it to the LASSO implementation in `graphicalVAR`. Part 2 illustrates the advantages of Bayesian inference for networks by focusing on the issue of network comparisons. First, we outline the relevance of methods for network comparison and the reasons why popular approaches from the cross-sectional setting cannot easily be applied to longitudinal models. Next, we explain the novel testing approach and assess its properties in our second simulation study. After a brief empirical example, we conclude with a discussion of the implications of our work and avenues for future research.

Part 1: Bayesian Inference for Idiographic Network Models

Estimation of GVAR models focuses on two networks, namely, the temporal network of lagged associations between the variables and the contemporaneous network of the associations between the innovations. There are many possible ways to estimate such a model in a Bayesian framework. Here, we focus on a prior structure that assumes normal distributions for the regression parameters of the temporal network and places an appropriate matrix prior on the precision matrix of the contemporaneous network, which allows users to specify prior assumptions on the partial correlations of the innovations.

For our simulation studies, we rely on the implementation of Bayesian GVAR models in the `BGGM`-package by Williams et al. (2020). The `BGGM`-package implements Bayesian GVAR estimation by providing MCMC samples of the joint posterior distribution of the temporal and the contemporaneous network. We present a detailed, technical explanation of the Gibbs sampler in the supplementary material and instead focus on the model structure in the main text below. While `BGGM` is a user-friendly and fast software implementation, it offers a limited choice of priors. Specifically, the prior for the temporal coefficients must be centered around zero and assumes that the prior variance is identical for all parameters. Also, `BGGM` does not properly handle missing data, which also means that overnight effects cannot be removed (cf. Epskamp, 2020). As a remedy, we implemented the model and prior structure presented here in other software. Our R package tsnet provides an implementation of GVAR in the probabilistic programming language Stan (Stan Development Team, 2023) which provides more flexibility with respect to the priors and model setup, thereby addressing the limitations of BGGM. In a brief simulation study, we show that our Stan implementation results in virtually identical posterior point estimates as BGGM in the conditions that we studied. We additionally provide code to fit a Bayesian GVAR model in MPlus (Muthén & Muthén, 2017) in the online supplementary.

Standard GVAR models assume stationarity, meaning that means, variances, and covariances of all variables are constant over time. For reasons of parsimony, we restrict ourselves to a lag-1 structure for the temporal network. We standardize all variables, which is a default procedure in the literature on longitudinal network models (Bulteel et al., 2016).

In the following, $\boldsymbol{y}_t$ denotes the responses to p variables at time point t. Thus, $\boldsymbol{y}_{t-1}$ denotes the responses at the previous time point. All regression coefficients β_{ij} of lag-1 effects of variable j on variable i in the temporal network are collected in the $p \times p$ matrix $\boldsymbol{B}$. Moreover, $\boldsymbol{\zeta}_t$ is a vector of normally distributed innovations with covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{\zeta}_t \tag{1}$$

$$\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2}$$

The inverse of the innovation covariance matrix provides the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. For the contemporaneous network, the partial correlations ρ_{ij} of the innovations of variables i and j are obtained from the off-diagonal elements of the precision matrix (Williams et al., 2020),

$$\rho_{ij} = \frac{-\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}. \tag{3}$$

We denote the partial correlation matrix as $\boldsymbol{P}$. To estimate GVAR networks in the BGGM-package, the user has to specify two prior distributions. For the temporal network, independent normal distributions centered at zero are assumed as priors for the regression coefficients. The user only has to set the standard deviation s_β to an appropriate value:

$$\beta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, s_\beta). \tag{4}$$

Alternatively, one could choose specific priors for different parameters, which may be non-centered with different variances, in our Stan implementation.

Regarding the prior distribution for the contemporaneous network, the user specifies a scale parameter that determines the expected size of all partial correlations (Williams et al., 2020). This follows from the precision matrix being approximately distributed as $IW(\delta + p - 1, \boldsymbol{I}_p)$, with p being the number of variables, shape parameter δ, and $\boldsymbol{I}_p$ a $p \times p$ identity matrix (Williams et al., 2020). This allows us to define a (marginal) prior directly on the partial correlations, specifically, a scaled beta distribution on the interval from -1 to $+1$:

$$\rho_{ij} \sim \text{Beta}\left(\frac{\delta}{2}, \frac{\delta}{2}\right) \ \text{ scaled to } \ [-1, 1] \tag{5}$$

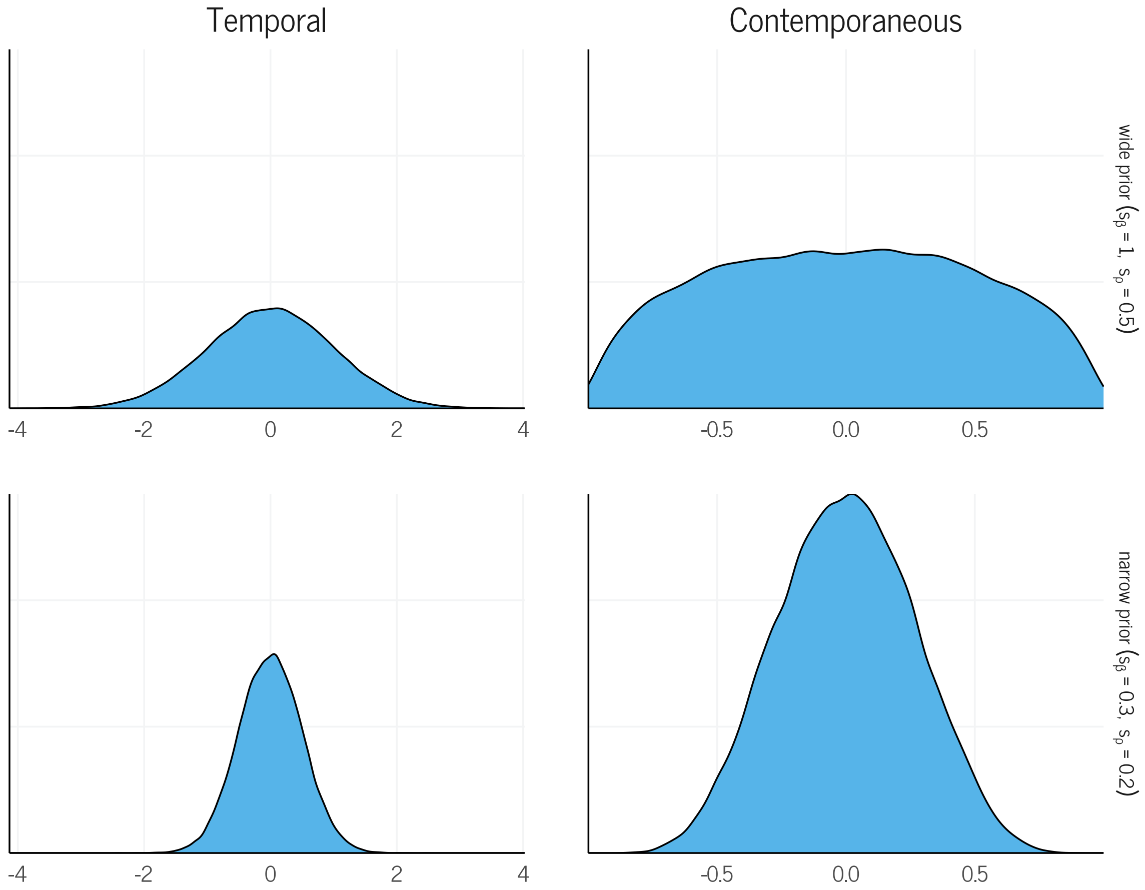
The standard deviation of this prior distribution is $s_\rho = \frac{1}{\delta+1}$. Hence, the user can define a prior by (a) defining a plausible value for the standard deviation s_ρ of non-zero partial correlations and (b) specifying δ such that $\delta = (s_\rho)^{-1} - 1$.[3]

We illustrate the priors for GVAR models in Figure 1 by showing two possible combinations of prior hyperparameters that are used in the remainder of the manuscript. The plots show the density of 50,000 samples drawn either from a more diffuse prior ($s_\rho = 0.5$ and $s_\beta = 1$) or from a more informed prior ($s_\rho = 0.3$ and $s_\beta = 0.2$). Under the wider prior, a substantial proportion of sampled VAR parameters of the temporal network

[3] In the BGGM package, a Matrix-F prior distribution is used for the precision matrix, which is a flexible and computationally convenient prior distribution since it is conditionally conjugate to the precision matrix. It is a scale mixture of a Wishart and an inverse-Wishart distribution (IW; Williams & Mulder, 2020). In `BGGM`, the prior hyperparameters of the Matrix-F prior are fixed to specific values such that the precision matrix is approximately defined with an Inverse Wishart distribution. The user only has to specify the prior hyperparameter s_ρ to define the prior standard deviation of the partial correlations. For more details on the matrix F-prior, see the supplementary material and Williams and Mulder (2020).

Figure 1

Illustration of the GVAR model with various prior distributions.



Note. The plot shows the density of 50,000 prior samples of regression coefficients in the temporal network (left panels) and of partial correlations in the contemporaneous network (right panels).

are greater than 1 and many partial correlations of the contemporaneous network are greater than 0.5. This may be considered implausible for standardized variables in practical applications, as indicated by previous simulation designs (Hoekstra et al., 2022; Mansueto et al., 2023). Setting a more informed prior (i.e., a narrower distribution) can have an effect similar to ridge regression in the frequentist setting, where parameter estimates are shrunk towards zero without setting the parameters strictly to zero (van Erp et al., 2019).

We now turn to a simulation-based comparison of Bayesian and LASSO estimation of GVAR models to evaluate the performance of both methods in different settings. In the following, we report only a brief summary of the simulation results and provide more details in the supplement, where we follow the reporting guidelines of Siepe, Bartoš, et al. (2023).

Simulation 1: Performance of Bayesian Estimation

Methods

Software and Setup. We used R version 4.3.2 (R Core Team, 2023), `BGGM` version 2.1.0 (Williams & Mulder, 2021), and `graphicalVAR` version 0.3 (Epskamp & Asena, 2021; Epskamp, Waldorp, et al., 2018) for all simulations and analyses. R code for reproducing the analyses, detailed session information, and supplementary materials are available at the Open Science Framework (https://osf.io/9byaj/). Data were simulated using the `graphicalVARsim` function from the `graphicalVAR`-package. All variables were standardized. We used 1,000 repetitions for each simulation condition.

Data Generation. We generated data under 5 (sample size) × 4 (data-generating processes) = 20 simulation conditions. We used five different sample sizes $n \in \{50, 100, 200, 400, 1000\}$.

We investigate performance in networks with different structures by using an empirical sparse network, a simulated sparse network, and a simulated non-sparse network as data-generating processes. We focus on networks with six or eight variables as previous simulations have shown inadequate performance in larger networks with typical psychological data of only a few hundred observations at most (Mansueto et al., 2023). Investigating the performance of dense graphs is important because it is plausible that true edges are never exactly zero, and because network estimates of prior research often used sparse data-generating processes only. A visualization of data-generating networks is provided in Appendix 1.

Estimation with graphicalVAR. We relied on network estimation as implemented in the `graphicalVAR`-package, a frequently used R package providing state-of-the-art methods. The model for the temporal and contemporaneous network is defined above in Equations (1) and (2). A more detailed description of the implementation of LASSO GVAR is provided in the supplement and in Epskamp, Waldorp, et al. (2018). We additionally estimated unregularized graphicalVAR models by setting the LASSO shrinkage parameters λ_B and λ_Θ to 0 for both networks.

Estimation with BGGM. We used the function `var_estimate` in the `BGGM`-package to estimate Bayesian idiographic networks as explained above. We used a grid of different priors ranging from informative to rather diffuse priors. In the manuscript, we show results for a wider prior setting (the `BGGM` default of $s_\beta = 1$ and $s_\rho = 0.5$) and a narrower prior setting ($s_\beta = 0.2$ and $s_\rho = 0.3$).

We focus on a Bayesian modeling approach that does not perform structure selection (i.e., it does not set coefficients to zero). As a remedy, we additionally used thresholding based on credible intervals (CI). This approach sets edge estimates to 0 if the corresponding CI contains 0, a strategy that has previously been used for cross-sectional and longitudinal networks (Burger et al., 2022; Jongerling et al., 2022; Williams, 2021).

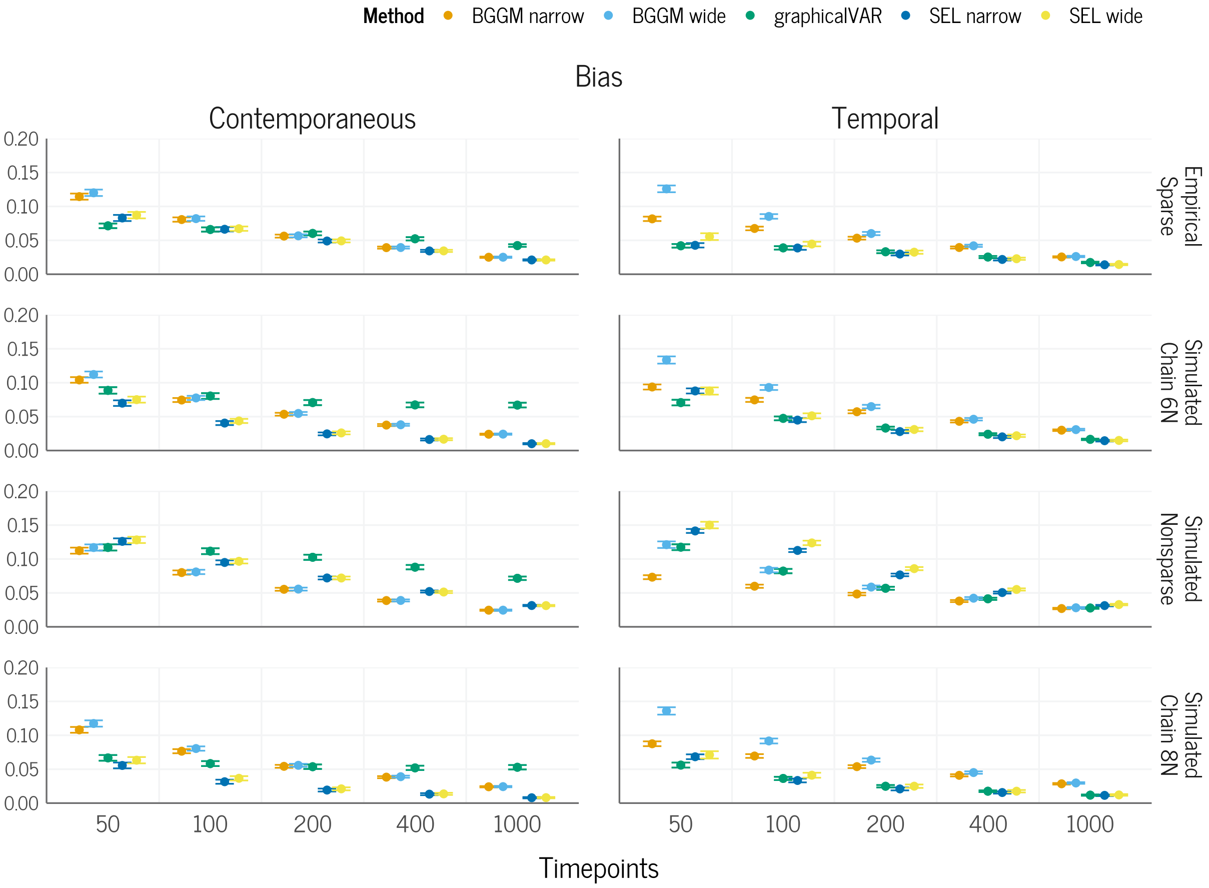
Performance Metrics. We based our choice of performance metrics on previous simulation studies investigating the performance of `graphicalVAR` (Mansueto et al., 2023). We calculated bias and correlation with the true parameters for all estimation methods. For the estimation methods that set coefficients to zero, we further investigated sensitivity (true-positive rate) and specificity (true-negative rate). For the `BGGM` method, we also computed the coverage rate and the width of credible intervals.

Results

First, as expected, the performance of all methods depended strongly on the number of time points. With 50 time points, the examined methods achieved mediocre performance at best. Figure 2 shows that the performance of all methods became more similar with more time points. Second, the best-performing method in terms of (absolute) bias depended on the data-generating mechanism. Under sparsity, LASSO and thresholded Bayesian estimation performed best, while non-thresholded Bayesian GVAR showed advantages when the data-generating network was dense, outperforming both LASSO GVAR and its

Figure 2

Absolute Bias for Different Simulation Conditions.



Note. The two columns show the contemporaneous and temporal network while rows show the different data-generating processes. Estimation methods are shown in different colors in the same order as they appear in the legend. "BGGM" denotes Bayesian estimation without thresholding, "SEL" denotes Bayesian estimation with thresholding (i.e., edge selection). Horizontal lines indicate $\pm 1 \times$ SE.

unregularized version. Third, we found that narrower priors on β and a relatively wider prior on ρ seemed to perform best. Fourth, Bayesian thresholding was conservative and showed poor sensitivity to detect nonzero edges in smaller sample sizes, especially in the temporal network. Here, LASSO GVAR generally performed better. Note that the thresholding approach was not specifically designed for the task of reliable edge detection or structure selection (Sekulovski et al., 2023). Instead, thresholding based on credible intervals is a pragmatic ad hoc solution because it simply dichotomizes continuous posterior distributions of edges. Still, sensitivity can be increased by using smaller credible intervals (e.g., 80% CIs) with the disadvantage of achieving lower specificity. A plot illustrating this trade-off is provided in the supplementary material. Fifth, the credible intervals for Bayesian GVAR showed good frequentist coverage properties, with a median absolute difference of 0.6% between empirical coverage and credible interval width across all simulation conditions. This indicates that Bayesian uncertainty quantification is promising for GVAR modeling.

In summary, we found that LASSO performs well for estimating idiographic networks based on longitudinal data, especially under sparse data-generating processes. Bayesian GVAR also shows a good performance which was often comparable to or, depending on the data, sometimes even higher than that of LASSO. The results suggest that in sparse graphs, both thresholding and LASSO outperformed non-thresholding methods, with LASSO and thresholded Bayesian estimation with narrower priors performing roughly equally well in terms of bias. For a non-sparse graph, the continuous regularization of Bayesian estimation via priors outperformed thresholding and LASSO methods in terms of bias and correlation. Overall, the results show that the match between the structure of the data-generating process and the type of estimation method determines performance. If true networks are dense, which is a plausible assumption given that edges may often not be perfectly zero, the regularization by LASSO may not always be the best option available.

Part 2: Testing Differences Between Idiographic Networks

The idea of estimating person-specific network models is driven by the assumption that heterogeneity in person-specific processes matters in areas such as psychopathology (Bringmann et al., 2013). It may be tempting for researchers to focus on visual displays of estimated networks to determine whether individuals differ (e.g., regarding the structure of temporal associations between symptoms). However, as shown in previous simulation studies (Hoekstra et al., 2022) and in the first part of the present manuscript, reliable estimation of GVAR models requires a large number of observations. Commonly available psychological time-series data provide only relatively few time points which may not be sufficiently informative, in turn leading to unstable results (Mansueto et al., 2023). Hoekstra et al. (2022) showed that examining only point estimates of different idiographic network structures can lead to a false appearance of heterogeneity due to sampling variability and a lack of statistical power. As a remedy, new methods are required to assess whether idiographic networks that 'look different' (in terms of point estimates) result from actual, true differences between individuals or are merely different due to sampling variability (Hoekstra et al., 2022). In the following, we briefly summarize the literature on comparison methods for cross-sectional networks. Next, we use Bayesian GVAR estimation to develop a new test that assesses whether data are sufficiently reliable to conclude that the data-generating processes underlying two network models actually differ.

Comparing networks across groups or individuals is of major interest in network science in general (Tantardini et al., 2019). In psychology, several approaches allow researchers to compare networks estimated on cross-sectional data between two or more groups (Haslbeck, 2022). Cross-sectional comparisons rely on methods such as random permutation of group membership (van Borkulo et al., 2022), moderation analyses using group membership as a predictor (Haslbeck, 2022), extensions of gLASSO for multiple groups (Costantini et al., 2021), several Bayesian approaches (Williams et al., 2020), and traditional significance testing (Haslbeck, 2022). Haslbeck (2022) provides an overview of these methods and their performance.

While several methods for comparisons of cross-sectional networks are available, it is not straightforward to transfer these approaches to longitudinal, time-series data (Hoekstra et al., 2022). For example, both the network comparison test (van Borkulo et al., 2022) and Bayesian methods (Williams et al., 2020) rely on merging data of different groups to create a reference model under the assumption that the groups are equal. Such a strategy cannot easily be applied to time-series data since the order of observations matters and because data from different individuals cannot simply be combined into a single data set. Therefore, our goal was to develop a new approach for comparing idiographic networks based on time-series data.

The proposed comparison method uses the full posterior distribution to account for uncertainty in network estimation. Similar to various cross-sectional approaches, it is a 'global test' as it tests the hypothesis that two network models share the same data-generating process (i.e., $\boldsymbol{B}_a = \boldsymbol{B}_b$ and $\boldsymbol{\Theta}_a = \boldsymbol{\Theta}_b$) without focusing on the detection of differences for specific edges. Suppose that two GVAR models are estimated for individuals a and b and that we want to compare the corresponding temporal networks (note that the procedure is identical for the contemporaneous networks). In estimating the GVAR model for two individuals, we do not set any edges to zero to keep the full information about the uncertainty of all parameter estimates. Our goal is to determine whether the data provide enough evidence that differences in estimated edges between the two fitted models are not just due to sampling variability.[4]

To compare estimated networks, it is necessary to quantify the discrepancy between a large number of parameters. For each network, parameter estimates are collected in two matrices where $\boldsymbol{B}$ represents the temporal network and $\boldsymbol{P}$ the partial correlation matrix of the contemporaneous network. In a first step, we compute two matrices that contain the differences of all parameters:

$$\boldsymbol{D}_B = \boldsymbol{B}_a - \boldsymbol{B}_b \tag{6}$$

$$\boldsymbol{D}_P = \boldsymbol{P}_a - \boldsymbol{P}_b. \tag{7}$$

To quantify the differences in all parameter estimates with a single number, we compute a norm of all elements of the difference matrix $\boldsymbol{D}$. In linear algebra, norms describe the magnitude or size of a vector or matrix. Norms are ubiquitous in statistics and are used in many applications such as LASSO (Tibshirani, 1996), network comparison methods (Tantardini et al., 2019; Ulitzsch et al., 2023), or in change detection for time series

[4] As time series data are usually standardized for GVAR models, we do not consider differences in intercepts.

(Cabrieto et al., 2018). For our purpose, we chose three norms that are applied to the vector of all elements of the difference matrix $\boldsymbol{D}$. Specifically, we implement the Frobenius norm (i.e., the Euclidian or ℓ_2-norm of the vector of all matrix elements),

$$\|\boldsymbol{D}\|_2 = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} D_{ij}^2},$$

the absolute-value norm (ℓ_1-norm of the vectorized matrix),

$$\|\boldsymbol{D}\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} |D_{ij}|,$$

and the maximum norm (ℓ_∞-norm of the vectorized matrix),

$$\|\boldsymbol{D}\|_{\max} = \max_{i,j\in\{1,\ldots,N\}} |D_{ij}|.$$

By computing the norm of the difference matrices of the parameter estimates (i.e., posterior means), we obtain the estimated distance $\hat{d}$ that describes the estimated discrepancy between the networks for data sets a and b. We focus on the difference between posterior means as these would be the numerical estimates typically plotted and reported by applied researchers. However, to judge whether an observed distance is relatively large compared to sampling variability, we need a reference distribution. This reference should reflect the uncertainty in parameter estimates under the null hypothesis that there are no true differences in the data-generating parameters. To create a reference distribution, we randomly draw $R = 1,000$ pairs of samples (i.e., the two sets of parameter values in iterations s_1 and s_2) from the posterior samples of model A without replacement. Using posterior samples to compute transformed quantities of interest (e.g., differences in parameters) is a common technique in Bayesian modeling, for example, in posterior-predictive checks (Berkhof et al., 2000). Usually, transformed quantities are computed separately for each iteration during MCMC sampling (i.e., for a single, possible multivariate vector of parameter values). Our approach deviates from typical Bayesian approaches in that we draw *pairs* of samples from the same posterior distribution (e.g., the posterior samples $\boldsymbol{B}_a^{(s_1)}$ and $\boldsymbol{B}_a^{(s_2)}$ from iterations s_1 and s_2) to calculate the difference matrix for each pair. By drawing two samples from the *same* posterior distribution (i.e., that for data set a), the reference distribution only reflects the amount of estimation uncertainty while assuming the same underlying data-generating process. To account for potential problems due to a high autocorrelation of samples, we draw pairs that are sufficiently far apart in the MCMC chain so that they can be considered independent.[5] We then compute the norm for all posterior-sampled difference matrices to obtain a reference distribution for the observed norm.

To compare the empirical distance between networks a and b against the reference distribution, we assess whether the estimated distance $\hat{d}$ is greater than a certain proportion of posterior distances (95% by default). If this is the case, we conclude that the data provide sufficient evidence that two networks actually differ with respect to the underlying data-generating processes. Since we are generating two reference distributions (i.e., one for a and one for b), it is necessary to use a decision rule on how to aggregate the results of the two comparisons. After initial simulations, we decided that a test result is considered to be positive when at least one of the two comparisons indicates a difference in networks (we refer to this as 'OR-rule' below). We chose this rule to increase the sensitivity of the test for detecting differences between networks. Importantly, the proposed test can only indicate whether differences in parameter estimates are larger than can be expected by mere sampling noise. However, the test cannot provide evidence *for* the null hypothesis, meaning that we cannot conclude that two networks are identical because this requires a specific hypothesis about the expected amount of discrepancies (see Discussion).

To summarize, the proposed test for the comparison of temporal networks requires

[5] This is achieved by splitting the posterior samples into two halves based on the iteration index, and then randomly drawing pairs from the two halves. More details are explained in the manual of the `tsnet` R package.

the following steps:

1. Estimate separate Bayesian GVAR models for data sets a and b to obtain $s = 1, \dots, S$ posterior samples $\boldsymbol{B}_a^{(s)}$ and $\boldsymbol{B}_b^{(s)}$.

2. Compute the empirical distance between the point estimates (i.e., posterior means) of the two data sets using a specific norm: $\hat{d} = \|\hat{\boldsymbol{B}}_a - \hat{\boldsymbol{B}}_b\|$

3. Separately for each data set, randomly assign $r = 1, \dots, R$ pairs of posterior samples (each indexed by s_1 and s_2) and compute the difference matrix for each pair: $\boldsymbol{D}_a^{(r)} = \boldsymbol{B}_a^{(s_1)} - \boldsymbol{B}_a^{(s_2)}$ and $\boldsymbol{D}_b^{(r)} = \boldsymbol{B}_b^{(s_1)} - \boldsymbol{B}_b^{(s_2)}$

4. Compute the norm for all difference matrices of pairs of posterior samples: $d_a^{(r)} = \|\boldsymbol{D}_a^{(r)}\|$ and $d_b^{(r)} = \|\boldsymbol{D}_b^{(r)}\|$

5. Compute posterior-based p-values as the proportion of posterior samples for the norm that are larger than the estimated norm: $p_a = P\left(d_a^{(r)} > \hat{d}\right)$ and $p_b = P\left(d_b^{(r)} > \hat{d}\right)$

6. If at least one of the two p-values p_a or p_b is smaller than a certain criterion (we chose 5% as default), conclude that it is unlikely that the two data sets were generated by the same underlying process.

The same procedure applies to the contemporaneous network by using the partial correlation matrix $\boldsymbol{P}$ instead of $\boldsymbol{B}$ in all steps.[6] We implemented the test in the `tsnet` package in R (https://github.com/bsiepe/tsnet) along with further functionality such as model fitting in Stan and matrix posterior plots to visualize uncertainty when reporting results.

Simulation 2: Performance of the New Comparison Method

The second simulation study assesses the performance of the proposed test in a variety of settings. Specifically, we were interested in the power to detect true differences, and in the proportion of false-positives when the true, data-generating networks are identical.

[6] The cross-sectional comparison approach by Williams et al. (2020) uses the normalized precision matrix instead of the partial correlation matrix, which just has the reverse sign of the partial correlations.

Methods

Data Generation. We used the same six-node networks as data-generating processes as in Simulation 1. To manipulate the distance between two true networks, we created data-generating processes that differed by a certain degree from the original, data-generating process. For this purpose, we either changed one or multiple elements of the regression weights of the temporal network and the precision matrix of the contemporaneous network. Our approach resembles similar methods used in the cross-sectional literature (Haslbeck, 2022; van Borkulo et al., 2022; Williams et al., 2020).

Specifically, we implemented three qualitatively different ways of inducing differences in the true network. First, we changed the largest edge of both the $\boldsymbol{B}$ and $\boldsymbol{P}$ matrix by a factor of $\in \{1.4, 1.6\}$.[7] As a second approach, we added or subtracted a constant value (i.e., either 0.05, 0.1, or 0.15) from all elements of the original matrix $\boldsymbol{B}$ and $\boldsymbol{P}$. Whether the value was subtracted or added was decided randomly until the resulting matrix fulfilled certain criteria relevant to the convergence of models such as positive semi-definiteness (for details, see electronic supplement). Third, we permuted the order of variables, such that the column indices of the $p = 6$ variables were rearranged to $1, 3, 4, 2, 5, 6$. We added this condition to keep the absolute size of parameters identical between the original and the modified network. In total, this results in six modifications with different effect sizes. Changes of $\boldsymbol{P}$ were achieved by first changing elements of $\boldsymbol{\Theta}$, which were then scaled with respect to its diagonal elements to achieve a comparable effect on partial correlations for all data-generating processes.

Network Estimation. We used Bayesian GVAR estimation similarly as in Simulation 1. Here, we only used the more diffuse prior ($s_\rho = 0.5$ and $\sigma_\beta = 1$) as well as a more informed prior ($s_\rho = 0.3$ and $\sigma_\beta = 0.2$). We used the latter, narrower prior for the main results, and the wider prior for sensitivity analyses.

[7] We initially also used a factor of 1.2, but results were very similar to using a factor of 1.4, which is why we omitted this condition here.

Performance Metrics. In each simulation condition, we computed 1,000 pairwise network comparisons. To evaluate the performance of the test, we calculate the power to detect true differences and the proportion of false-positives for different simulation conditions.

Results

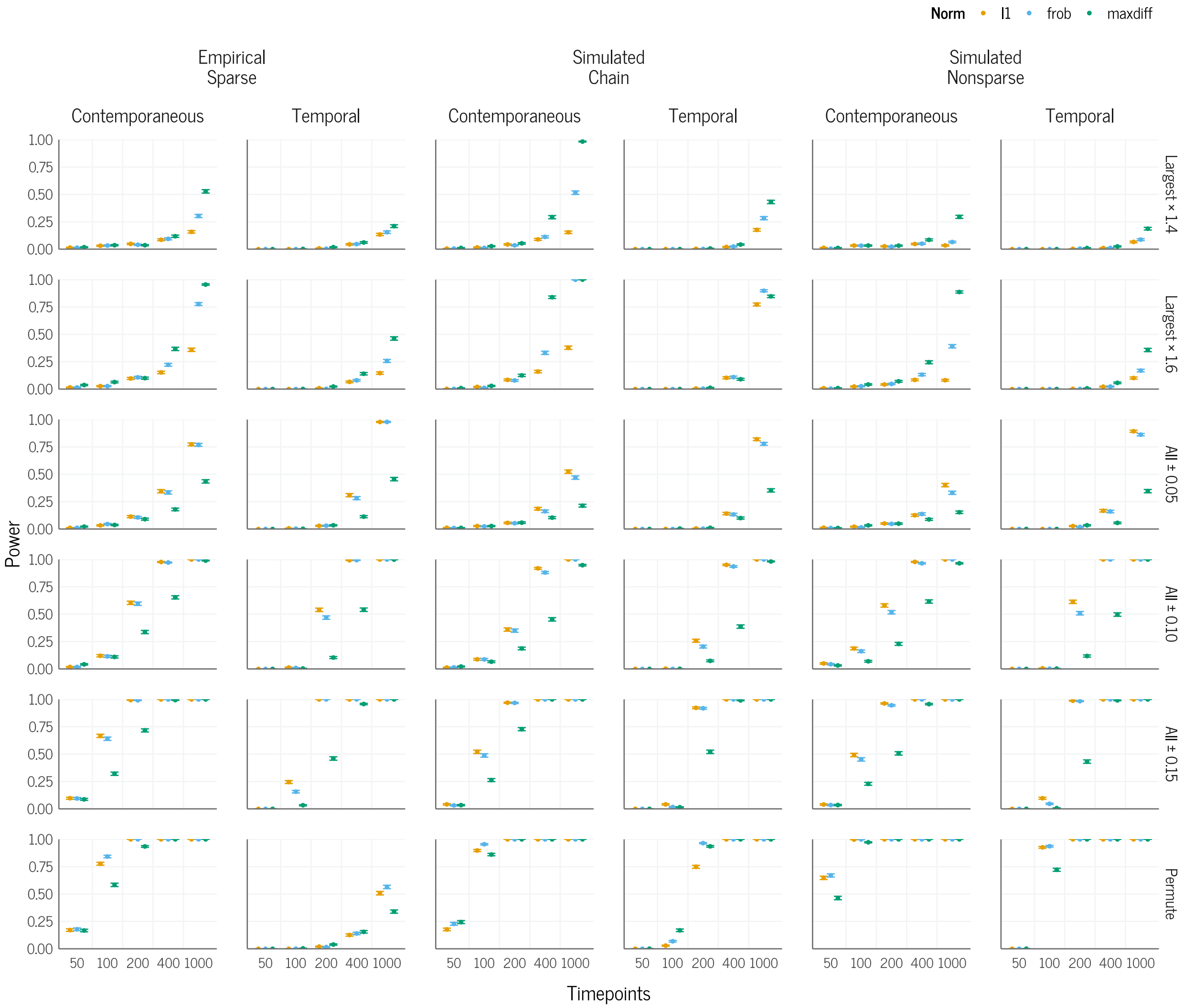
Figure 3 shows the power of the test (y-axis) with different numbers of time points (x-axis) across all manipulation conditions (rows) and data-generating processes (columns). Comparing the different rows shows that, the greater the difference between data-generating processes, the higher the power of the test. In the first two conditions, only a single edge changed, and accordingly, the power to detect this difference was small overall. This was different in the noise and permutation conditions, where a power of $> .80$ could often be achieved with 200 time points or even less. The power increased above .90 for larger sample sizes. Speaking of sample size, the power to detect differences was generally very low with only 50 time points, with power below .50 for all conditions except for some permutation conditions.

A comparison of the columns in Figure 3 does not show a clear pattern regarding the performance of the test for different manipulations of the data-generating process. Additionally, it is unclear whether the test generally performs better for the contemporaneous or the temporal network, since results differed strongly depending on the data-generating process. A comparison of the different norms shows that the maximum norm performs best when only one edge is changed. In other conditions, the other two norms worked better, with the Frobenius norm being slightly better than the ℓ_1-norm in some conditions.

Figure 4 shows the proportion of false positives with a gray horizontal line at the nominal level of 5% used by the comparison test. In all conditions, the proportion of false positives was below the nominal value in smaller sample sizes. For the temporal networks, the proportion of false-positive came closer to the nominal value for larger sample sizes. This was not the case for the contemporaneous networks.

Figure 3

Power of the Comparison Test for Idiographic Networks.



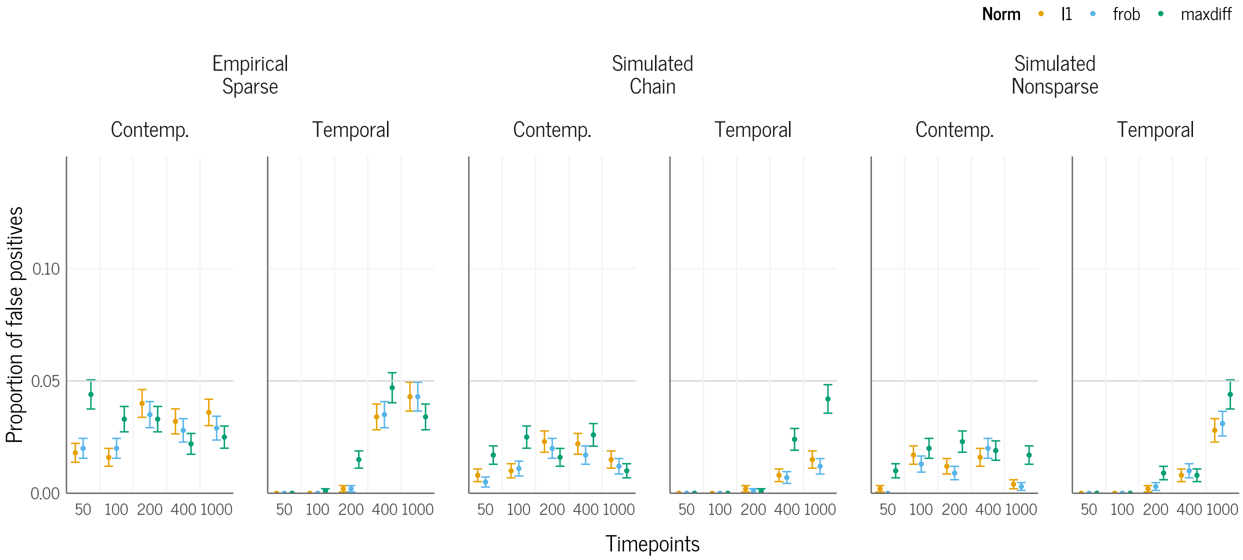
Note. Rows show different manipulations of the true, data-generating network. Power of the test to detect differences in idiographic networks with a prior of $s_\rho = 0.3$ and $\sigma_\beta = 0.2$. Vertical bars indicate $\pm 1 \times SE$, not adjusted for boundedness of the scale. Norms from left to right: ℓ_1-norms in orange, Frobenius norm in light blue, and maximum norm in green.

Prior Sensitivity

Figures 5 and 6 show prior sensitivity analyses for selected simulation conditions. Across all conditions, the wider prior led to a higher power to detect differences as well as false-positive rates closer to nominal rates. However, the wide prior was too liberal in small sample sizes in some conditions. A narrower prior led to more conservative results (i.e., less

Figure 4

False-Positive Rate of the Test.



Note. Vertical bars indicate 1 × SE, not adjusted for boundedness of the scale. Norms from left to right: ℓ_1-norms in orange, Frobenius norm in light blue, and maximum norm in green.

evidence for the presence of differences), especially for smaller sample sizes. An inspection of the sampling distribution of probability values of the test under the null (i.e., when assuming that the two data-generating processes are identical) is provided in the supplement.

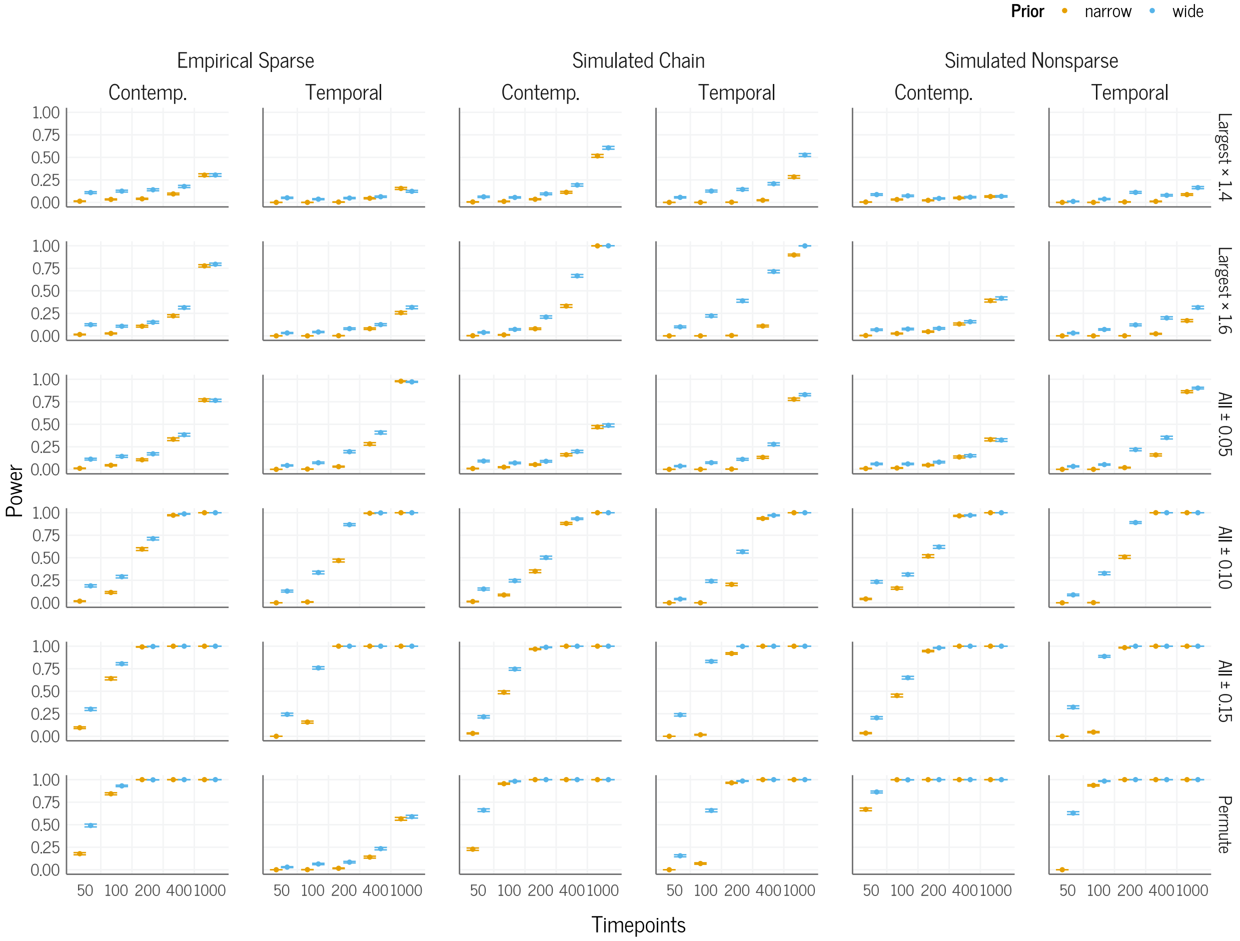
Summary

The results show that, as expected, power increased for more time points and stronger manipulations (i.e., larger effect sizes). Furthermore, comparisons of temporal networks had higher power than those of contemporaneous networks. False-positive rates were close to the nominal value for a wider prior, while a narrower prior led to more conservative results overall. The narrower prior generally leads to more regularization which is beneficial for estimation. However, regularization also has the effect that estimated networks become more similar to each other, making it harder to detect differences between them.

Regarding the assumed effect size for manipulating the true networks, it is currently unclear which amount of differences between networks is realistic or practically relevant. Whether differences in a few, specific edges or in all edges are plausible depends on the type

Figure 5

Prior Sensitivity of the Power of the Test.

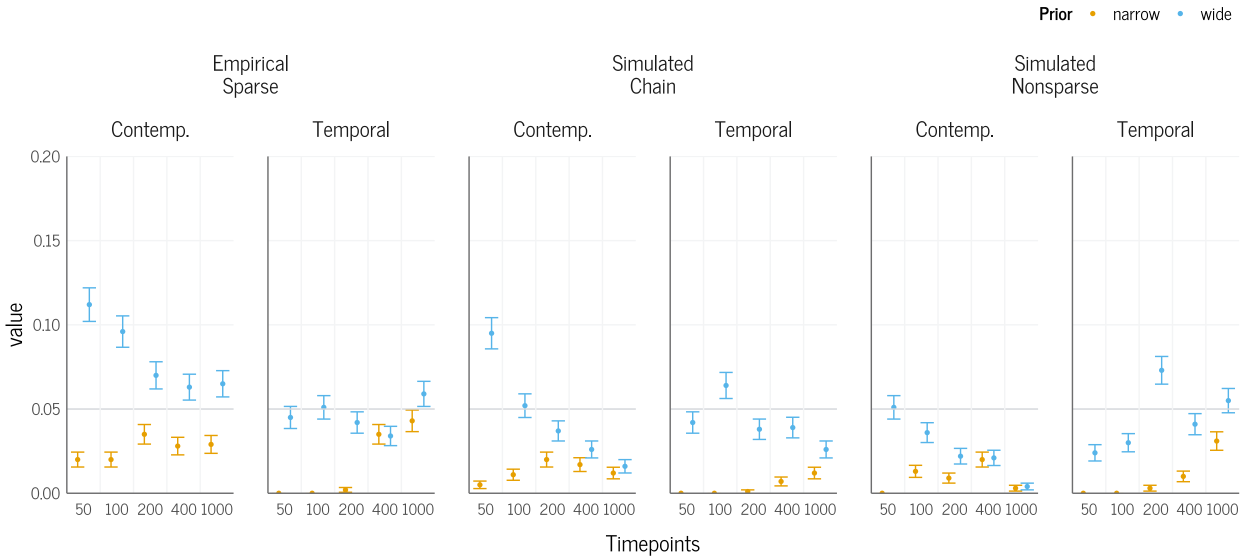


Note. Power of the comparison test in idiographic networks using only the Frobenius Norm. Narrow prior left (orange), wide prior right (light blue). Vertical bars indicate $\pm 1 \times SE$, not adjusted for boundedness of the scale.

of variables included in the model and on the research question. Also, the performance of the test and the magnitude of plausible differences between networks depends on the size of the network, which we did not investigate further here. Regardless, the test presented here is conservative and errs on the side of caution. If necessary, it can be made less conservative by changing the decision threshold to less than 5%. Regarding the choice of a specific norm, our simulations did not provide a clear picture that would allow us to provide general recommendations. Because differences in all parameter values, such as those created in the noise and permutation conditions, seem more plausible than differences in just one edge, we

Figure 6

Prior Sensitivity of the False-Positive Rate of the Test



Note. False-positive rate of the comparison test in idiographic networks using only the Frobenius Norm. Narrow prior left (orange), wide prior right (light blue). Vertical bars indicate ±1 × SE, not adjusted for boundedness of the scale.

tentatively recommend using the Frobenius norm.

Empirical Example

We now turn to an empirical example to illustrate Bayesian estimation of idiographic networks using `BGGM` and the new comparison test using `tsnet`. The empirical example also serves as a plausibility check for the manipulation conditions in our second simulation study. While we induced increasing differences between data-generating processes that can easily be interpreted (e.g., changing the largest value), the amount and type of differences between models in empirical applications are unknown. If we found no differences at all in this example, our test may be underpowered for empirical applications. Below, we present a comparison of two individuals for which we did not find evidence for actual differences, although estimated networks may appear qualitatively different. We also compare all possible pairs of individuals in the sample against each other.

We use data previously analyzed by Fisher et al. (2017) which are available at the Open Science Framework (https://osf.io/5ybxt/). Details about the design and data collection can be found in Fisher and Boswell (2016). The sample that we use consists of data by 40 individuals with either Major Depressive Disorder or Generalized Anxiety Disorder. Participants were asked to complete four daily surveys sent to their smartphones for at least 30 days before receiving psychotherapy. In these surveys, participants rated their current symptoms, affect, behavioral avoidance, and reassurance seeking (21 items in total) on a visual analogue scale from 0 to 100. The mean number of available surveys per individual after pre-processing was 133.2.

Data Preprocessing. We selected six variables (content, fatigue, concentrate, positive, hopeless, enthusiastic) based on the individual and sample-aggregated marginal distributions of responses, prioritizing small floor effects and approximately normal distributions, if possible. We then followed the pre-processing steps of Fisher et al. (2017). First, we removed linear trends in the data by regressing each item for each individual on the timestamp and replacing the raw responses with the residuals of this regression. Cubic spline interpolation was then applied to the data to account for the different time intervals between measurement occasions occurring due to nighttime. This approach is further described and evaluated in a small simulation in Fisher et al. (2017), while Epskamp, van Borkulo, et al. (2018) provide evidence that this approach is likely inferior to removing overnight effects (which is not possible in BGGM).

Network Estimation. We estimated Bayesian GVAR models using the BGGM package. We set the prior hyperparameters $s_\rho = 0.25$ and $s_\beta = 0.5$ and visualized the resulting networks using the qgraph package (Epskamp et al., 2012). We chose the hyperparameters for the empirical example based on our simulation results, with slightly narrower prior for both networks for better estimation performance, without being overly conservative for detecting differences in networks. For prior-sensitivity analyses, we further used prior hyperparameters $s_\rho \in \{0.1, 0.5\}$ and $s_\beta \in \{0.25, 1\}$. For the comparison test, we used the Frobenius norm and set the decision threshold to 5% while also exploring other norms for sensitivity analyses. Convergence diagnostics of MCMC sampling are implemented in BGGM and tsnet and shown in the supplement.

Uncertainty Visualizations

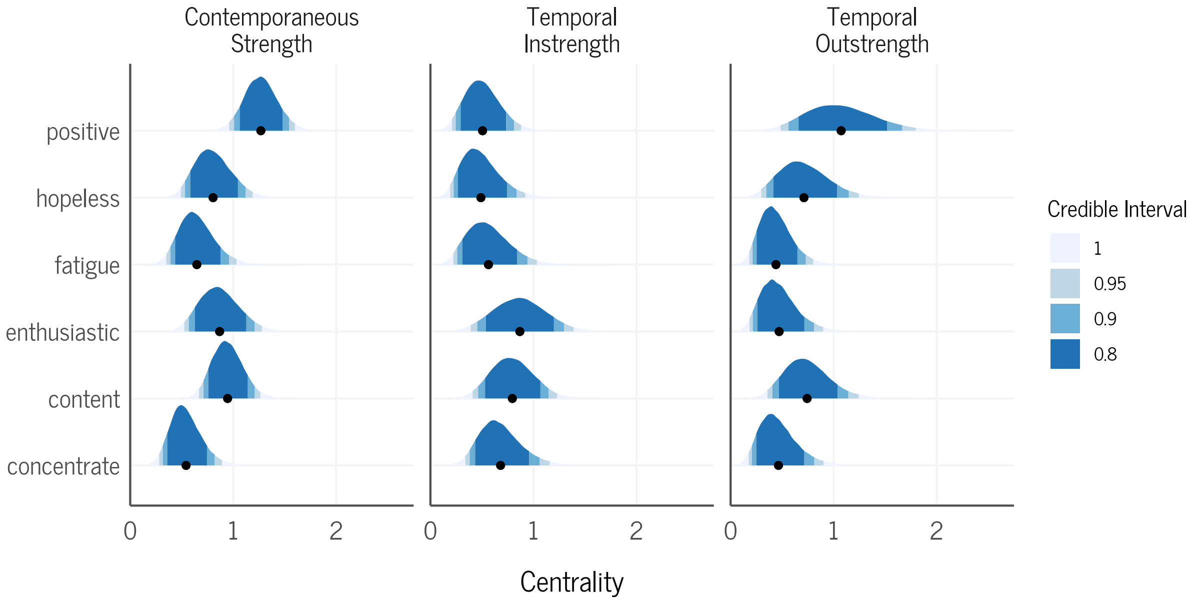
One of the main advantages of Bayesian inference is the assessment and visualization of estimation uncertainty. Besides quantifying the uncertainty for each edge via the (marginal) posterior distribution, we can use Bayes Factors or other Bayesian approaches to test hypotheses regarding certain edges. For the example, focusing on the temporal association between $positive_{t-1}$ and $enthusiastic_t$ of an example participant (ID 12) with a posterior mean of 0.312 (95% CI: $[0.034, 0.587]$), we can compute a Bayes Factor to conduct a two-sided test of this edge against zero. As explained in Oravecz and Vandekerckhove (2023), testing temporal coefficients of a VAR model can be used to quantify evidence for and against Granger causality. [8] We obtain a Bayes Factor of 3.26, which indicates relatively weak support for the hypothesis of the edge being nonzero compared to the null hypothesis, given our data and prior. However, when using Bayes factors, one should perform sensitivity analyses and rely on robust workflows (Heck et al., 2022; Schad et al., 2022), as illustrated in the online supplement. Prior sensitivity analyses in the supplement show that the relative evidence for the alternative hypothesis compared to the null hypothesis changes slightly for different prior choices, but all analyses tend to indicate weak evidence for the alternative hypothesis.

Second, as we alluded to in the introduction, Bayesian inference allows us to quantify the uncertainty in transformed quantities of GVAR parameters, such as network centrality measures, based on the posterior samples. As researchers are often interested in using such network summaries to interpret the results of these highly parameterized models (Bringmann et al., 2022), a Bayesian approach can be helpful to highlight the uncertainty in results. In Figure 7, we show the posterior distribution of three different network centrality measures for all edges: Temporal instrength (the sum of all absolute edges predicting a specific variable),

[8] More formally, we compare the models $\mathcal{M}_0 : \beta = 0$ and $\mathcal{M}_1 : \beta \sim \mathcal{N}(0, 0.5)$. We can then compute the Savage-Dickey ratio of the prior and posterior density at $\beta = 0$ to compute the Bayes factor BF_{10}.

Figure 7

Centrality Uncertainty for ID 12.



Note. This figure shows posterior distributions of three centrality indices for individual 12. Different credible interval sizes are colored with increasing brightness. The black dots represent posterior means.

temporal outstrength (the sum of all absolute temporal edges from a variable to others), and contemporaneous strength (the sum of all absolute contemporaneous edges of a variable). We can, for example, see that "positive" has the highest posterior mean in temporal outstrength, but there is considerable uncertainty around this estimate. We could use this information further and compute the difference between the posterior distributions of two centrality measures to quantify the evidence that one node is more central than another.

Results for Two Individuals

As an illustration, we present the results for two individuals for whom the comparison test did not indicate differences in either the temporal or the contemporaneous network. Figure 8 shows the estimated networks for participants 12 and 18 and a visualization of the comparison test. A visual inspection of the two estimated networks shows that they look quite similar overall, with some apparent differences in the strength of the edges, especially in the temporal network. The mean absolute edge difference between the networks was 0.112

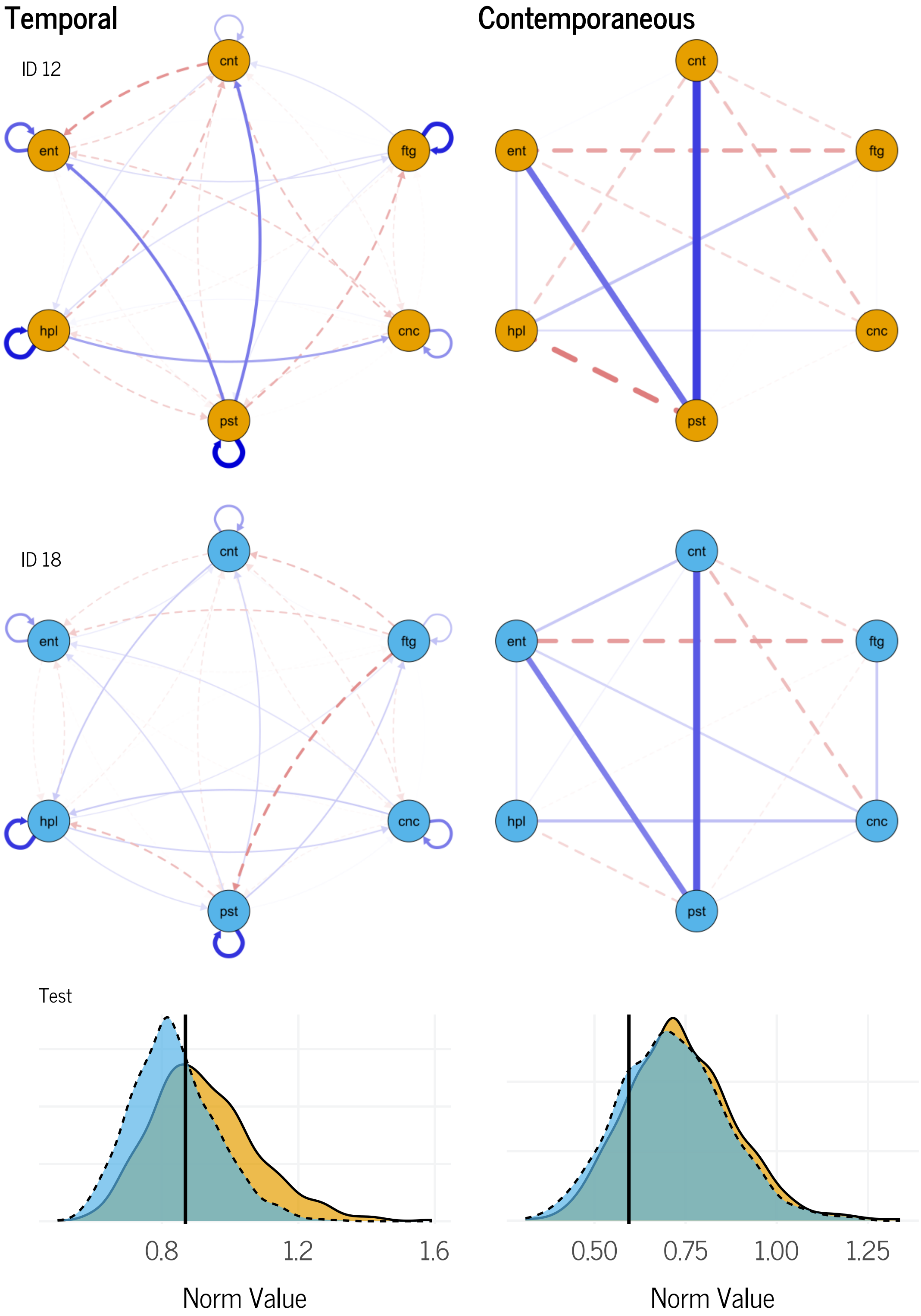
for the temporal and 0.085 for the contemporaneous, respectively. Additionally, there were 12 edges with a different sign in the temporal network and 4 in the contemporaneous network.

To illustrate the comparison test, the bottom row of Figure 8 shows that the empirical distance between the two networks (horizontal black line) was smaller than a substantial part of the reference distribution. Thus, we do not have sufficient evidence to conclude that the two networks arose from different data-generating processes. As the modeling function in `BGGM` does not automatically include sampling diagnostics or warnings if something goes wrong during MCMC sampling, it is crucial to check this manually. We show how to check convergence in the R code associated with this manuscript.

All Pairwise Comparisons. We also performed all $\frac{40!}{2!\,(40-2)!} = 780$ possible pairwise comparisons for the full sample. Using the Frobenius norm, we found evidence for differences in 87.4% of all comparisons for contemporaneous networks and in 38.2% of all comparisons for temporal networks.The comparison of all possible pairs of individuals shows four main points. First, we found evidence of differences in a substantial proportion of all comparisons, suggesting that our test may be sufficiently sensitive in practice. Second, the norms differed in how often the test indicated meaningful differences between persons. Using the Frobenius norm resulted in the largest number of positive results, followed by the absolute-value (ℓ_1-)norm and, at some distance, the maximum norm. Third, we also observed a higher proportion of positive test results for the contemporaneous network, which somewhat contradicts our simulation results. Possible explanations for this result include a larger heterogeneity in contemporaneous associations or a difference between empirical and simulated data, for example, because contemporaneous networks are more densely connected than simulated. Fourth, as we show in the supplement, the prior has a non-negligible effect on the number of positive test results, with narrower priors generally leading to more conservative results. Specifically, setting $s_\rho = 0.5$ and $s_\beta = 1$ resulted in evidence for differences in 91.92% (contemporaneous) and 51.79% (temporal) of comparisons. Using a narrow prior with hyperparameters $s_\rho = 0.1$ and $s_\beta = 0.25$ resulted in almost no differences

Figure 8

Illustration of Network Comparison for Two Individuals.



Note. Edges are scaled with respect to the maximum edge. None of the edges are set to zero, while the estimates for some edges are so small that they may be hard to see. Red (dashed) lines indicate negative edge weights, blue (solid) lines indicate positive ones. Vertical lines in the test panel indicate the empirical distance.

found. This is probably due to excessive shrinkage of all parameters towards zero which renders estimated networks more similar across participants. We did not correct for multiple testing in this example.

Discussion

We highlighted the benefits of Bayesian estimation with continuous regularization for idiographic network models and evaluated its performance against LASSO estimation. Moreover, we developed and evaluated a comparison test for detecting differences in idiographic network models and applied it in an empirical example.

Bayesian Inference and Assumptions about Sparsity

In Part 1, we showed that using LASSO works well overall for the estimation of idiographic networks in multiple contexts, especially in smaller sample sizes and when assuming a sparse ground truth. Yet, LASSO is not always the best option for achieving good performance. Bayesian GVAR performed similarly well as LASSO in many scenarios, especially when assuming a non-sparse ground truth. Choosing an appropriate method for estimating GVAR networks thus depends on the degree of sparsity assumed by the researcher. This assumption, sometimes also termed the *bet on sparsity* (Hastie et al., 2017), has been discussed extensively in the cross-sectional network literature. There are two main reasons for taking this bet. Statistically, LASSO regularization can decrease false-positives in models with many parameters and facilitate interpretation (Epskamp & Fried, 2018). Theoretically, assumptions about the degree of sparsity in network structures are closely linked to the debate about common-cause versus network theories (Epskamp et al., 2017). When the goal is to map all symptoms of a particular disorder or those of a combination of disorders from a network-theoretic standpoint, sparsity plays an important role in cross-sectional research. However, when assuming a common latent cause for the co-occurrence of symptoms, this results in a dense network, for which LASSO may not be an appropriate method (Epskamp et al., 2017).

In a longitudinal setting, using LASSO can be motivated by similar reasons.

Statistically, our results showed that LASSO performs well when estimating networks in small samples. However, there may be good reasons for researchers to divert from the assumption of sparsity in GVAR networks. First, contrary to the cross-sectional setting, idiographic networks are rarely used to map out the whole symptom network of a disorder or even the bridges between multiple disorders. Rather, they often include only a few substantively related variables that may be chosen for pragmatic or theoretical reasons (Bringmann et al., 2022). Further, a large-scale panel network study of transdiagnostic symptom associations resulted in a large dense network (O'Driscoll et al., 2022). Although not directly transferable to idiographic networks, these results still provide evidence for the plausibility of dense longitudinal networks. When assuming a dense ground truth, the method for network estimation should reflect this assumption (see Epskamp et al., 2017). This is not merely a theoretical issue, as our simulations showed that Bayesian GVAR without edge selection may outperform LASSO for dense networks.

Comparing Idiographic Network Models

In Part 2, we introduced a novel test to assess and detect differences between GVAR models. This test uses matrix norms to compare empirical differences between all estimated parameters while taking the uncertainty reflected by the posterior distribution into account. The new test proved to be conservative in our simulations while showing good false-positive rates. The test is not suited to detect differences in a specific edge but works best if differences occur across multiple edges. We tentatively recommend the use of the Frobenius norm for network comparisons (which corresponds to the Euclidian distance), but future work is necessary to evaluate other measures of differences in networks. In our empirical example, we found evidence for differences in networks for a sizable proportion of pairwise comparisons of individuals. We also implemented the novel test and other useful functionality in the `tsnet` package in R. Focusing on a matrix norm of the difference between two matrices provides a lot of flexibility for testing purposes. It could therefore also be used for other network models, both in temporal and in cross-sectional data (Ulitzsch et al., 2023). As the precision matrix is estimated to be saturated in our model, users can also only test for differences in the temporal matrix, in other words, the VAR model only. Furthermore, the implementation in `tsnet` allows to include only a subset of specific edges when computing the norm of matrix differences.

How should the new comparison test be applied in practice? In general, one may argue that we do not test a very interesting null hypothesis in the first place. Indeed, assuming that the data-generating process is perfectly identical for two individuals may often seem implausible. Nevertheless, a negative result of the comparison test is still informative. Such a result serves as a cautionary reminder that empirically observed differences in network estimates may merely be due to sampling variability. Put differently, there may be too much noise in the data to detect any true difference between individuals. This also implies that, based on the estimated networks only, choosing different, optimally tailored treatments for the two individuals may be premature.

Beyond inter-individual comparisons, the novel comparison test may prove even more useful for intra-individual comparisons. The method can be used to test the stability of idiographic networks over time. For example, an individual may provide daily self-report data a few weeks both before and after a treatment or a major life event. If long time series are available, they may be split in half to estimate separate networks for the first and second segments. The new comparison test may be applied to test whether the two resulting networks are different.

Issues in obtaining evidence for the null

A main conceptual limitation of the proposed comparison test concerns its inability to provide evidence for the null hypothesis. Negative test results can occur both due to large estimation uncertainty or due to the actual invariance of data-generating processes. In developing the test, we also tried out various approaches for obtaining evidence for the null hypothesis of network equality (e.g., computing a Bayes Factor or relying on regions of equivalence). However, the general setup of first fitting separate idiographic GVAR models for individuals and then comparing the results is, in our opinion, not suitable for computing Bayes factors or quantifying evidence for the null more broadly. By specifying independent prior distributions for all individuals, we implicitly define a certain prior on the expected differences between two networks. This prior distribution under the alternative hypothesis is very uninformative, meaning that we expect large differences between two networks. Using overly vague priors with Bayes Factors can lead to undesirable results (Heck et al., 2022) such as a strong tendency to favor the null hypothesis (Morey & Rouder, 2011). The issue of defining independent priors for two groups (Dablander et al., 2022) can be addressed by specifying a joint model for two (or more) individuals which includes an explicit prior distribution for the expected differences between networks.

Obtaining evidence for the null hypothesis generally requires the specification of precise expectations for the alternative hypothesis, both in Bayesian and frequentist (Neyman-Pearson) approaches to hypothesis testing. However, specifying prior assumptions for the expected discrepancy between two networks is not an easy task. In the context of idiographic networks, where we test for differences in many parameters simultaneously, it is difficult to specify prior expectations about theoretically plausible or practically relevant differences between networks. Gaining a better understanding of how time series networks are expected to differ is an important topic for future research. In the meantime, our test provides a first safeguard against an overinterpretation of differences between idiographic networks, which may lead to premature conclusions in favor of heterogeneity in the literature or practical applications.

Limitations and Future Research

While Bayesian inference improves the assessment of uncertainty, it does not solve one of the main problems of idiographic modeling in psychological data, namely, small numbers of observations per individual (Mansueto et al., 2023). Performance with sample sizes below 100 is generally weak regardless of the chosen method. Additional criticisms of discrete-time (G)VAR models include, among others, the assumption of stationarity (i.e., that model parameters are constant over time), the assumption of discrete time intervals (i.e., that time intervals between measurement occasions are equally spaced), and the assumption of error-free measurement (Bringmann, 2021; Haslbeck & Ryan, 2021; Ryan & Hamaker, 2021; Schuurman et al., 2015). Researchers who want to implement idiographic network models in clinical practice should carefully consider which conclusions can be drawn in realistic settings. For samples with many individuals but only a few time points, one may instead consider group-based methods such as multilevel VAR (Bringmann et al., 2013) or Group Iterative Multiple Model Estimation (GIMME; Beltz & Gates, 2017). These approaches pool information across individuals and likely perform better in many cases (Lafit et al., 2021). Instead of replacing these methods, our test complements the methodological toolbox for network modeling.

The present work provides a first evaluation of the performance of a Bayesian GVAR approach for idiographic networks. We focused on a specific model and prior setup which could be generalized in future work. Specifically, the topic of Bayesian variable selection could be explored further. To set coefficients to exactly zero with Bayesian GVAR, we used thresholding based on credible intervals (CIs). This approach performed well in terms of bias in several simulation conditions. However, the approach also lacked sensitivity for smaller sample sizes. The weak performance of CI-based thresholding is consistent with criticisms of this approach in the cross-sectional network literature (Sekulovski et al., 2023). The present manuscript focused mainly on quantifying estimation uncertainty of network parameters, a main advantage of Bayesian inference. Future work should also investigate more principled Bayesian approaches for assessing uncertainty about the overall network structure and the inclusion versus exclusion of edges based on the psychometric literature on cross-sectional network estimation (e.g., Marsman et al., 2022; Williams & Mulder, 2020) and on the econometric literature on Bayesian time series modeling (Ahelegbey et al., 2016; Giannone et al., 2015; Paci & Consonni, 2020). Bayesian model averaging could be used to average over all plausible structures to properly account for uncertainty about the network structure

(Marsman & Haslbeck, 2023).

Bayesian GVAR estimation also facilitates idiographic network modeling in clinical practice as one may implement edge-specific priors. Thereby, one can incorporate expert knowledge about pairwise associations between variables (Burger et al., 2022). While we did not explore such a strategy here, our Stan implementation allows researchers to implement edge-specific informative priors (Jongerling et al., 2022). As noted above, the application of GVAR models still suffers from limited precision in typical psychological data, so incorporating information from other individuals may be beneficial. While there are Bayesian methods for estimating multilevel models in longitudinal data (see, for example, Hamaker et al., 2018; Li et al., 2022), these could be improved with further work on prior choices and structure uncertainty.

Besides alternative Bayesian test approaches mentioned above, comparison tests for idiographic networks may also be constructed in different ways. First, data of different individuals could be chained together to create a reference model under the assumption of equality (see Park et al., 2022). We attempted to implement these ideas, but initial simulations did not show promising results. Second, simulation-based procedures similar to ours could be performed via parametric bootstrapping for frequentist models. Highlighting the relevance of the topic, during the revision of the present manuscript, Hoekstra et al. (2023) published a promising network comparison test based on ideas of invariance testing. We consider the evaluation of such alternative approaches and their comparison with the one presented here as worthwhile avenues for future research.

Conclusion

We presented Bayesian estimation for GVAR models and evaluated its performance in a simulation study. Our simulations showed that a Bayesian GVAR approach without edge or structure selection provides a viable alternative to using LASSO regularization in certain conditions, especially if the underlying network structure is dense. We also developed and evaluated a new test to compare idiographic network models across individuals. The test is conservative and may serve as a safeguard against premature conclusions about the presence of true heterogeneity. Overall, Bayesian inference for longitudinal network models allows researchers to assess the estimation uncertainty of idiographic networks in a principled way.

References

Ahelegbey, D. F., Billio, M., & Casarin, R. (2016). Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics*, *31*(2), 357–386. Retrieved April 25, 2023, from https://www.jstor.org/stable/26609615

Beck, E. D., & Jackson, J. J. (2020). Consistency and change in idiographic personality: A longitudinal ESM network study. *Journal of Personality and Social Psychology*, *118*(5), 1080–1100. https://doi.org/10.1037/pspp0000249

Beltz, A. M., & Gates, K. M. (2017). Network Mapping with GIMME. *Multivariate behavioral research*, *52*(6), 789–804. https://doi.org/10.1080/00273171.2017.1373014

Berkhof, J., van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*(3), 337–354. https://doi.org/10.1007/s001800000038

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

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 1–18. https://doi.org/10.1038/s43586-021-00055-w

Bringmann, L. F. (2021). Person-specific networks in psychopathology: Past, present, and future. *Current Opinion in Psychology*, *41*, 59–64. https://doi.org/10.1016/j.copsyc.2021.03.004

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy*, *149*, 104011. https://doi.org/10.1016/j.brat.2021.104011

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0060188

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using Raw VAR Regression Coefficients to Build Networks can be Misleading. *Multivariate Behavioral Research*, *51*(2-3), 330–344. https://doi.org/10.1080/00273171.2016.1150151

Burger, J., Epskamp, S., van der Veen, D. C., Dablander, F., Schoevers, R. A., Fried, E. I., & Riese, H. (2022). A clinical PREMISE for personalized models: Toward a formal integration of case formulations and statistical networks. *Journal of Psychopathology and Clinical Science*, *131*(8), 906–916. https://doi.org/10.1037/abn0000779

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., & Ceulemans, E. (2018). Testing for the presence of correlation changes in a multivariate time series: A permutation based approach. *Scientific Reports*, *8*(1), 769. https://doi.org/10.1038/s41598-017-19067-2

Costantini, G., Kappelmann, N., & Epskamp, S. (2021, February 10). *EstimateGroupNetwork: Perform the Joint Graphical Lasso and Selects Tuning Parameters* (Version 0.3.1). Retrieved April 6, 2023, from https://CRAN.R-project.org/package=EstimateGroupNetwork

Dablander, F., Huth, K., Gronau, Q. F., Etz, A., & Wagenmakers, E.-J. (2022). A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions. *Statistics in Medicine*, *41*(8), 1319–1333. https://doi.org/10.1002/sim.9278

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*, *85*(1), 206–231. https://doi.org/10.1007/s11336-020-09697-3

Epskamp, S., & Asena, E. (2021, October 19). *graphicalVAR: Graphical VAR for Experience Sampling Data* (Version 0.3). Retrieved March 29, 2023, from https://CRAN.R-project.org/package=graphicalVAR

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

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, *48*, 1–18. https://doi.org/10.18637/jss.v048.i04

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

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PLOS ONE*, *12*(6), e0179891. https://doi.org/10.1371/journal.pone.0179891

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. J. (2018). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science*, *6*(3), 416–427. https://doi.org/10.1177/2167702617744325

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate behavioral research*, *53*(4), 453–480.

Fisher, A. J., & Boswell, J. F. (2016). Enhancing the personalization of psychotherapy with dynamic assessment and modeling. *Assessment*, *23*(4), 496–506. https://doi.org/10.1177/1073191116638735

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of abnormal psychology*, *126*(8), 1044. https://doi.org/10.1037/abn0000311

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for Vector Autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451. https://doi.org/10.1162/REST_a_00483

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the cogito study. *Multivariate Behavioral Research*, *53*(6), 820–841. https://doi.org/10.1080/00273171.2018.1446819

Hamaker, E. (2012, January 1). Why researchers should think "within-person": A paradigmatic rationale. In M. Mehl & T. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Haslbeck, J. M. B. (2022). Estimating group differences in network models using moderation analysis. *Behavior Research Methods*, *54*(1), 522–540. https://doi.org/10.3758/s13428-021-01637-y

Haslbeck, J. M. B., & Ryan, O. (2021). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, 1–32. https://doi.org/10.1080/00273171.2021.1896353

Hastie, T., Friedman, J. H., & Tibshirani, R. (2017). *The elements of statistical learning : Data mining, inference, and prediction* (Second Edition, corrected at 12th printing). Springer.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., … Hoijtink, H. (2022). A review of applications of the Bayes factor in psychological research. *Psychological Methods*, *28*(3), 558–579. https://doi.org/10.1037/met0000454

Hoekstra, R. H. A., Epskamp, S., & Borsboom, D. (2022). Heterogeneity in individual network analysis: Reality or illusion? *Multivariate Behavioral Research*, *58*(4), 762–786. https://doi.org/10.1080/00273171.2022.2128020

Hoekstra, R. H. A., Epskamp, S., Nierenberg, A. A., Borsboom, D., & McNally, R. J. (2023, September 5). *Testing similarity in longitudinal networks: The Individual Network Invariance Test (INIT)*. https://doi.org/10.31234/osf.io/ugs2r

Jongerling, J., Epskamp, S., & Williams, D. R. (2022). Bayesian uncertainty estimation for gaussian graphical models and centrality indices. *Multivariate Behavioral Research*, 1–29. https://doi.org/10.1080/00273171.2021.1978054

Lafit, G., Meers, K., & Ceulemans, E. (2021). A systematic study into the factors that affect the predictive accuracy of multilevel VAR(1) models. *Psychometrika*. https://doi.org/10.1007/s11336-021-09803-z

Levinson, C. A., Hunt, R. A., Christian, C., Williams, B. M., Keshishian, A. C., Vanzhula, I. A., & Ralph-Nearman, C. (2022). Longitudinal group and individual networks of eating disorder symptoms in individuals diagnosed with an eating disorder. *Journal of Psychopathology and Clinical Science*, *131*(1), 58–72. https://doi.org/10.1037/abn0000727

Li, Y., Wood, J., Ji, L., Chow, S.-M., & Oravecz, Z. (2022). Fitting Multilevel Vector Autoregressive Models in Stan, JAGS, and Mplus. *Structural equation modeling : a multidisciplinary journal*, *29*(3), 452–475. https://doi.org/10.1080/10705511.2021.1911657

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

Marsman, M., Huth, K., Waldorp, L. J., & Ntzoufras, I. (2022). Objective bayesian edge screening and structure selection for ising networks. *Psychometrika*, *87*(1), 47–82. https://doi.org/10.1007/s11336-022-09848-8

Marsman, M., & Haslbeck, J. M. (2023, March 3). *Bayesian analysis of the ordinal markov random field*. https://doi.org/10.31234/osf.io/ukwrf

Marsman, M., & Rhemtulla, M. (2022). Guest Editors' Introduction to The Special Issue "Network Psychometrics in Action": Methodological Innovations Inspired by Empirical Problems. *Psychometrika*, *87*(1), 1–11. https://doi.org/10.1007/s11336-022-09861-x

Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*, 201–218. https://doi.org/10.1207/s15366359mea0204_1

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

Muthén, L., & Muthén, B. (2017, April). *Mplus User's Guide* (Version Eight Edition). Los Angeles, US.

O'Driscoll, C., Epskamp, S., Fried, E. I., Saunders, R., Cardoso, A., Stott, J., Wheatley, J., Cirkovic, M., Naqvi, S. A., Buckman, J. E. J., & Pilling, S. (2022). Transdiagnostic symptom dynamics during psychotherapy. *Scientific Reports*, *12*(1), 10881. https://doi.org/10.1038/s41598-022-14901-8

Oravecz, Z., & Vandekerckhove, J. (2023). Quantifying evidence for—and against—granger causality with Bayes factors. *Multivariate Behavioral Research*, *0*(0), 1–11. https://doi.org/10.1080/00273171.2023.2214890

Paci, L., & Consonni, G. (2020). Structural learning of contemporaneous dependencies in graphical VAR models. *Computational Statistics & Data Analysis*, *144*, 106880. https://doi.org/10.1016/j.csda.2019.106880

Park, J. J., Chow, S.-M., Epskamp, S., & Molenaar, P. (2022, October 15). *Subgrouping with Chain Graphical VAR Models*. https://doi.org/10.31234/osf.io/u3ve8

Piccirillo, M. L., Beck, E. D., & Rodebaugh, T. L. (2019). A clinician's primer for idiographic research: Considerations and recommendations. *Behavior Therapy*, *50*(5), 938–951. https://doi.org/10.1016/j.beth.2019.02.002

Ryan, O., & Hamaker, E. L. (2021). Time to intervene: A continuous-time approach to network analysis and centrality. *Psychometrika*. https://doi.org/10.1007/s11336-021-09767-0

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of bayes factors. *Psychological Methods*. https://doi.org/10.1037/met0000472

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology*, *6*, 1038. https://doi.org/10.3389/fpsyg.2015.01038

Sekulovski, N., Keetelaar, S., Huth, K., Wagenmakers, E.-J., Bork, R. van, Bergh, D. van den, & Marsman, M. (2023, April 19). *Testing conditional independence in psychometric networks: An analysis of three bayesian methods.* https://doi.org/10.31234/osf.io/ch7a2

Siepe, B. S., Bartoš, F., Morris, T., Boulesteix, A.-L., Heck, D. W., & Pawel, S. (2023, October 31). *Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting.* https://doi.org/10.31234/osf.io/ufgy6

Siepe, B. S., Kloft, M., & Heck, D. W. (2023, July 19). *Bayesian Estimation and Comparison of Idiographic Network Models.* https://doi.org/10.31234/osf.io/uwfjc

Siepe, B. S., Kloft, M., & Heck, D. W. (2024, January 31). Online Supplementary Material for Bayesian Estimation and Comparison of Idiographic Network Models. *Open Science Framework*. https://osf.io/9byaj/

Stan Development Team. (2023). *Stan Modeling Language Users Guide and Reference Manual* (Version 2.33). https://mc-stan.org

Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, *9*, 17557. https://doi.org/10.1038/s41598-019-53708-y

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Ulitzsch, E., Khanna, S., Rhemtulla, M., & Domingue, B. W. (2023). A graph theory based similarity metric enables comparison of subpopulation psychometric networks. *Psychological Methods*. https://doi.org/10.1037/met0000625

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1–26. https://doi.org/10.1038/s43586-020-00001-2

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217. https://doi.org/10.1037/met0000100

van Borkulo, C. D., van Bork, R., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2022). Comparing network structures on three aspects: A permutation test. *Psychological Methods*. https://doi.org/10.1037/met0000476

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Williams, D. R. (2021). Bayesian estimation for gaussian graphical models: Structure learning, predictability, and network comparisons. *Multivariate Behavioral Research*, *56*(2), 336–352. https://doi.org/10.1080/00273171.2021.1894412

Williams, D. R., & Mulder, J. (2020). Bayesian hypothesis testing for Gaussian graphical models: Conditional independence and order constraints. *Journal of Mathematical Psychology*, *99*, 102441. https://doi.org/10.1016/j.jmp.2020.102441

Williams, D. R., & Mulder, J. (2021, August 20). *BGGM: Bayesian Gaussian Graphical Models* (Version 2.0.4). Retrieved March 29, 2023, from https://CRAN.R-project.org/package=BGGM

Williams, D. R., Rast, P., Pericchi, L. R., & Mulder, J. (2020). Comparing gaussian graphical models with the posterior predictive distribution and Bayesian model selection. *Psychological Methods*, *25*(5), 653–672. https://doi.org/10.1037/met0000254

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

Ye, A., Gates, K. M., Henry, T. R., & Luo, L. (2021). Path and directionality discovery in individual dynamic models: A regularized unified structural equation modeling approach for hybrid vector autoregression. *Psychometrika*, *86*(2), 404–441. https://doi.org/10.1007/s11336-021-09753-6

Appendix 1

The following plots display the data-generating matrices we used to generate data. Temporal directed networks are on the left, and contemporaneous undirected networks are on the right. Directed edges between nodes in a temporal network indicate lag-1 cross-lagged associations, whereas directed edges from a node to itself indicate autoregressive effects. In the contemporaneous network, undirected edges represent residual partial correlations. The thickness of an edge is scaled with respect to the size of the coefficient.

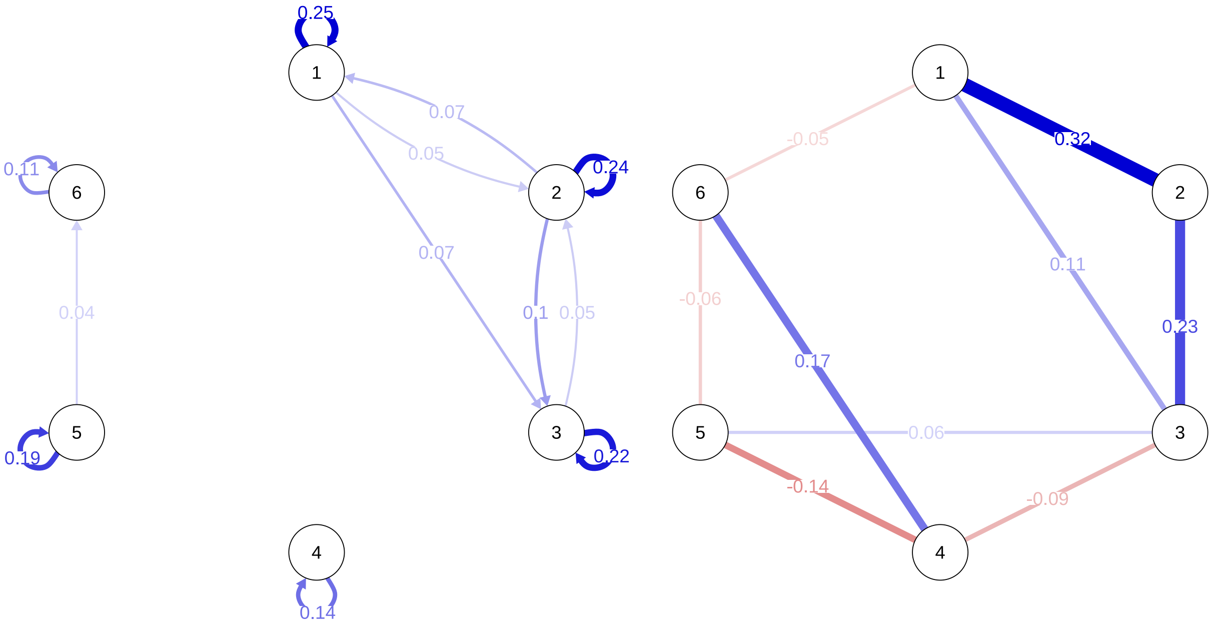


Figure 9

Empirical Sparse Network

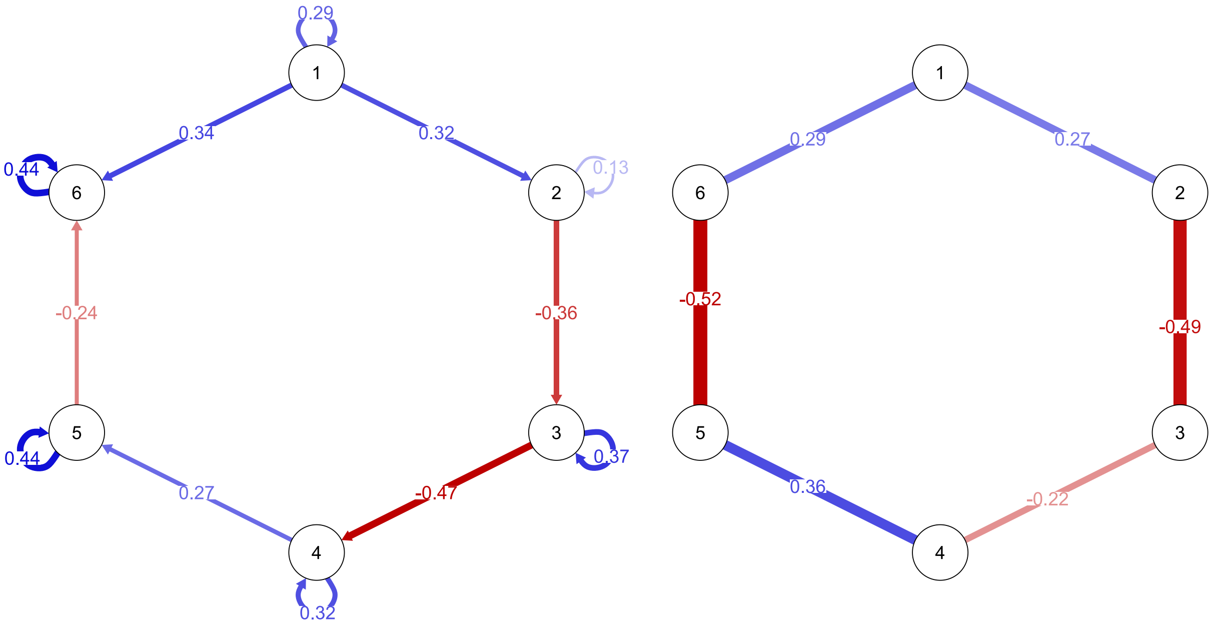


Figure 10

Sparse Chain with 6 Nodes

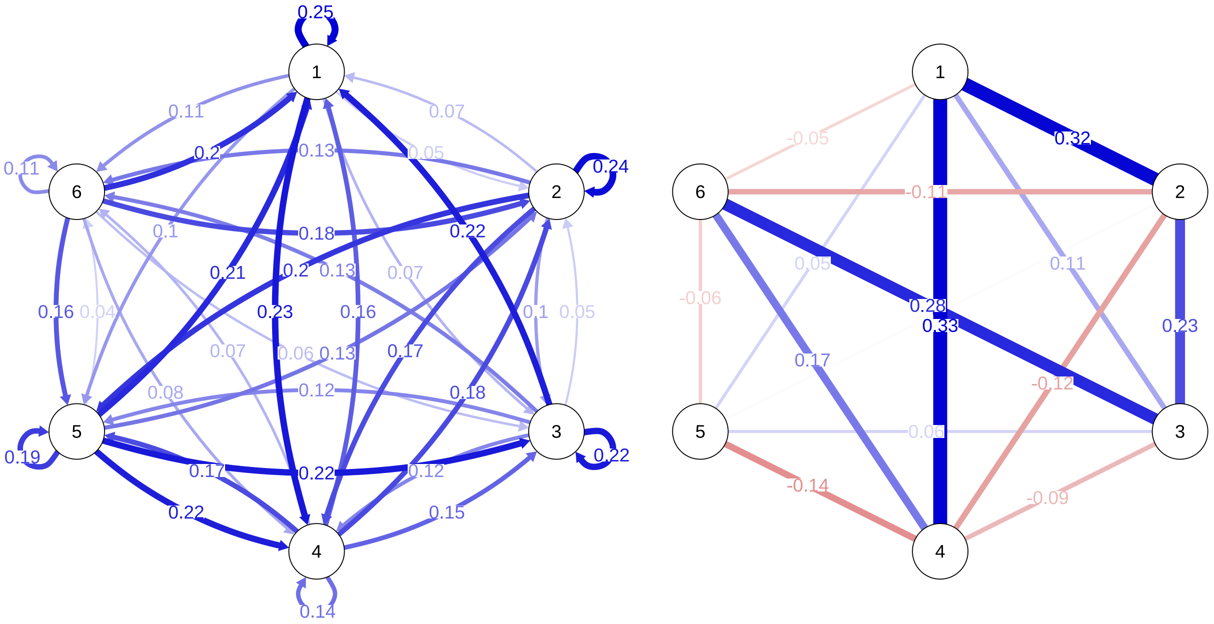


Figure 11

Nonsparse Network

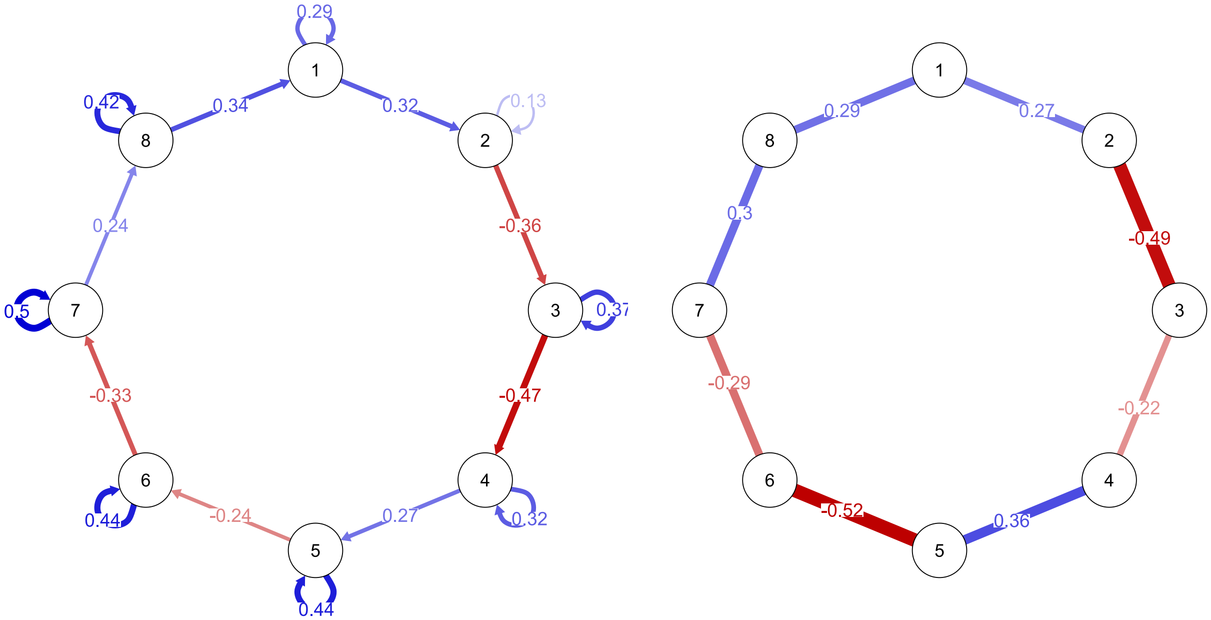


Figure 12

Sparse Chain with 9 Nodes