Bayesian Estimation and Comparison of Idiographic Network Models

Björn S. Siepe[1] and Daniel W. Heck[1]

[1]University of Marburg

Author Note

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

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; 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/.

We thank Maarten Marsman for helpful discussions on our approach, Raphael Hartmann and Matthias Kloft 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, Department of Psychological Methods, 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 such graphical vector autoregressive (gVAR) models uses LASSO regularization to estimate a contemporaneous network 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 first compare Bayesian and LASSO approaches across a range of conditions and performance measures in a simulation study. Overall, LASSO estimation performed well, while Bayesian gVAR may perform better when the true network is dense. 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. In a simulation study, the test was conservative and showed good false-positive rates. Finally, we apply Bayesian estimation and the novel testing approach in an empirical example using daily data on clinical symptoms for 40 individuals. Overall, Bayesian gVAR modeling facilitates the assessment of estimation uncertainty which is important for studying inter-individual differences of intra-individual dynamics.

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 symptoms has gained popularity as an alternative to classical conceptualizations of psychopathology (Borsboom, 2017). Alongside theoretical debates, the network perspective has inspired a number of methodological innovations and has found application in many areas of psychology (Borsboom et al., 2021). In typical psychological networks, *nodes* represent variables and *edges* statistical associations between nodes. Much of the early work in network modelling in psychology focused on cross-sectional data (e.g., Epskamp, Borsboom, et al., 2018). However, the interest in intra-individual dynamic associations between variables and inter-individual differences in the resulting network structures has led to increased use of longitudinal (or *dynamic*) network modelling, facilitated by the ease of data collection using mobile devices.

Theoretical and empirical work has 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 associations between variables *within* an individual *over time*. While multilevel or group network models are available for longitudinal data, there is a growing interest in purely idiographic approaches. These idiographic approaches are particularly important in clinical research due to their potential to provide a tailored perspective on an individual case and their alignment with the highly individualized perspective in clinical practice (Piccirillo et al., 2019). Idiographic network modelling has been used, for example, to investigate personality (Beck & Jackson, 2020), to inform treatment planning in eating disorders (Levinson et al., 2021), or to provide graphical feedback to clinicians (Hall et al., 2022).

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.[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) uses the residuals of the VAR model at each time point to estimate a *contemporaneous* network, which is intended to capture undirected effects between variables that occur faster than the lag interval of the temporal network. 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 jointly estimate both network structures and to shrink small coefficients to zero, as explained in detail below. 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. Note that there are also other approaches to obtain personalized networks, such as forms of structural VAR modelling (Epskamp, van Borkulo, et al., 2018; Ye et al., 2021), but we focus on the gVAR approach here due to its popularity and its use of regularization.

LASSO regularization can improve interpretability and increase specificity of estimated networks. However, it also has certain 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 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 6; Mansueto et al., 2020). The benefits of using LASSO regularization for such models with relatively few parameters are unclear.

[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.

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, results of idiographic network analysis may often appear to be more heterogeneous than they actually are. Especially 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 present the Bayesian gVAR approach, a Bayesian alternative to LASSO which accounts for the uncertainty in estimated idiographic networks. The approach uses Gibbs sampling as implemented by Williams in the *BGGM*-package (Williams & Mulder, 2021) to obtain the posterior distribution of all parameters in the temporal and the contemporaneous network. To the best of our knowledge, the performance of this approach has not yet been evaluated in the literature, nor has it been applied to empirical data. In this paper, we introduce the model and prior assumptions underlying the Bayesian approach and investigate its performance compared to LASSO in a simulation study. As a remedy for the issues outlined by Hoekstra et al. (2022), we use samples from the posterior distribution to test whether differences between two estimated networks indeed reflect genuine differences between individuals and not only mere sampling variability. Comparisons of idiographic networks are of interest both between individuals (e.g., Levinson et al., 2022) and within individuals over time (e.g., Beck & Jackson, 2020) to understand inter-individual differences as well as intra-individual stability of networks, respectively.

Hence, developing a test of network differences that is applicable to both types of settings fills an important gap in the literature. By harnessing the strengths of Bayesian estimation, we hope to make idiographic network estimation more robust and facilitate an assessment of uncertainty in the modeling process.

In the following, we briefly highlight advantages of Bayesian inference relevant to network analysis. However, we do not aim at providing 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 quantified in detail 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 implementation of idiographic networks in the *BGGM*-package. In a simulation study, we evaluate the performance of the Bayesian estimation method in different settings and

compare it to the 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 of 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

The *BGGM*-package by Williams et al. (2020) implements Bayesian gVAR estimation by providing MCMC samples of the full posterior distribution of the temporal and the contemporaneous network. We present a detailed, technical explanation of the Gibbs sampler in Appendix A and instead focus on the model structure in the following. Standard gVAR models assume stationarity, meaning that mean, variance, and covariance 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, $\boldsymbol{\epsilon}_t$ is a vector of normally distributed innovations (i.e., residual errors) with covariance matrix $\boldsymbol{\Sigma}$.

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

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

The inverse of the residual covariance matrix provides the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. For the contemporaneous network, the partial correlations ρ_{ij} of the residuals of variables i and j

are obtained from the elements of the precision matrix (Williams et al., 2020),

$$\rho_{ij} = \frac{-\boldsymbol{\Theta}_{ij}}{\sqrt{\boldsymbol{\Theta}_{ii}\boldsymbol{\Theta}_{jj}}}. \tag{3}$$

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:

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

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). The specification of the prior follows from first using a Matrix-F prior for the precision matrix (Mulder & Pericchi, 2018):[2]

$$\boldsymbol{\Theta} \sim F(\nu, \delta, \boldsymbol{C}), \tag{5}$$

with degrees of freedom ν, shape parameter δ, and scale matrix $\boldsymbol{C}$. The matrix-F prior is defined as a scale mixture of a Wishart and an inverse-Wishart distribution (IW; Williams & Mulder, 2020). The matrix-F distribution provides a flexible and computationally convenient prior distribution since it is conditionally conjugate to the precision matrix. Usually, however, we are not interested in the precision parameters. Instead, it is more intuitive to think about the implied partial correlations. The approach by Williams and Mulder (2020) solves this issue by allowing us to define a (marginal) prior directly on the partial

[2] Deviating from the original notation, we use $\boldsymbol{C}$ instead of $\boldsymbol{B}$ to denote the scale matrix of the matrix-F prior to avoid confusion with the matrix of regression weights $\boldsymbol{B}$.

correlations, specifically, a beta distribution on the interval from -1 to $+1$:

$$\rho \sim \text{Beta}\left(\frac{\delta}{2}, \frac{\delta}{2}\right) \text{ scaled to } [-1, 1] \tag{6}$$

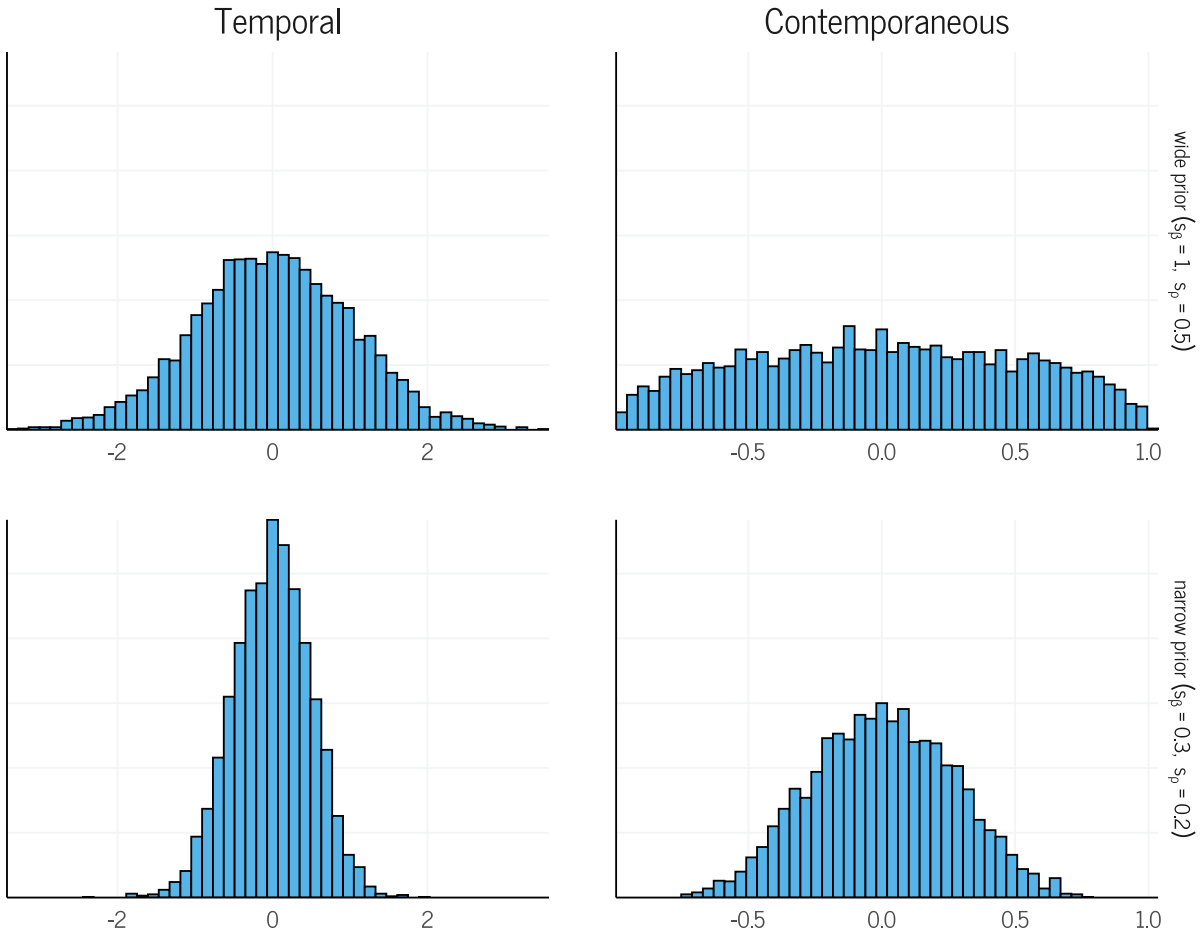
The standard deviation of this prior is $s_\rho = \frac{1}{\delta+1}$. Hence, the user can define a prior by plugging in an expected value for the standard deviation of non-zero partial correlations by choosing δ such that $\delta = (s_\rho)^{-1} - 1$. The other prior hyperparameters ν and $\boldsymbol{B}$ are fixed to specific values such that the precision matrix is approximately distributed as $IW(\delta + p - 1, \boldsymbol{I}_p)$, with p being the number of variables and $\boldsymbol{I}_p$ a $p \times p$ identity matrix. Hence, only the prior hyperparameter s_ρ has to be set by the user. Specifying a small number for the degrees of freedom and the identity matrix as the scale matrix results in a relatively uninformative prior (see Schuurman et al., 2016). For more details on the matrix F-prior, see Appendix A and Williams and Mulder (2020).

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 histograms show the distributions of 5,000 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 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., 2020). 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).

After defining priors and providing the data, the package BGGM returns posterior samples for the temporal and the contemporaneous network as well as (marginal) posterior means as point estimates. The results can be plotted using suitable packages in R (e.g.,

Figure 1

Illustration of different prior hyperparameters for gVAR models.



Note. The histograms show prior samples of regression coefficients in the temporal network (left panels) and of partial correlations in the contemporaneous network (right panels).

BGGM itself or *tsnet*, as presented later). 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.

Simulation 1: Performance of Bayesian Estimation

Methods

Software and Setup. We used R version 4.2.3. (R Core Team, 2022), *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, a detailed session information as well as supplementary materials are available at the Open Science Framework (https://osf.io/9byaj/). If applicable, we follow the steps of the *When to Worry and How to Avoid the Misuse of Bayesian Statistics* (WAMBS) checklist (van de Schoot et al., 2020), either in the manuscript itself or in the supplement, to increase the transparency and robustness of our analyses.

Data were simulated using the `graphicalVARsim` function from the *graphicalVAR*-package. All variables were standardized. We used 1,000 iterations for each simulation condition. We used the *doRNG*-package to set reproducible seeds across simulations conditions; details are included in the supplementary materials. Monte Carlo standard errors were calculated using the formulas by Morris et al. (2019). Standard errors for correlations were calculated as $SE = \sqrt{1 - r^2}/\sqrt{n_{\text{rep}} - 2}$, where r was the estimated correlation for a condition and n_{rep} was the number of simulation repetitions.

Data Generation. We generated data under 5 (sample size) $\times$ 4 (data-generating processes) = 20 simulation conditions. We used five different sample sizes $n \in \{50, 100, 200, 400, 1000\}$. Note that we use the terms 'sample size' and 'number of time points' interchangeably. Similar to previous simulation studies, sample sizes range from relatively small numbers, which are common in longitudinal studies, to very large numbers which are only found in single-case studies so far (Mansueto et al., 2020).

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., 2020). The first data-generating condition used a six-node network estimated on the empirical data by Fried et al. (2021) similar to Mansueto et al. (2020). This data-generating process is called *Empirical Sparse* hereafter because 24/36 and 6/15 edges were originally estimated to be

zero for the temporal and contemporaneous network, respectively. Mean absolute non-zero edge sizes were 0.123 (temporal) and 0.165 (contemporaneous).

As a second data-generating process, we created a six-node and eight-node sparse *Simulated Chain Graph* following Hoekstra et al. (2022), because previous simulation studies have shown good performance of *graphicalVAR* with such data structures (Hoekstra et al., 2022). In a chain graph, each node is only connected to two neighboring nodes in both networks. Mean absolute non-zero edge sizes were 0.333 (temporal) and 0.358 (contemporaneous) for the six-node graph, and 0.343 for both networks of the eight-node graph.

Third, we created a non-sparse data-generating process by adding random numbers drawn from a uniform distribution with the maximum set to the largest coefficient in the estimated network to all zero-coefficients of the *Empirical Sparse* graph. This graph is referred to as *Simulated Nonsparse* throughout the manuscript. We added this last condition to investigate performance when not assuming a sparse network as ground truth, similar to previous investigations in the cross-sectional network literature (Epskamp et al., 2017), while keeping ourselves to realistic parameter sizes. Investigating the performance for 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. Mean absolute non-zero edge sizes were 0.145 (temporal) and 0.142 (contemporaneous). Plots of the networks of all data-generating processes including the size of all edges are available in Appendix B.

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 setup for the temporal and contemporaneous network is identical as described above in Equations (1) and (2). Based on previous work establishing the multivariate regression with covariance estimation (MRCE) algorithm (Rothman et al., 2010), *graphicalVAR* jointly estimates the temporal and contemporaneous coefficients using

cyclical-coordinate descent (Epskamp, Waldorp, et al., 2018). Separate LASSO regularization parameters λ_B and λ_Θ are placed on the temporal coefficient matrix $\boldsymbol{B}$ and the precision matrix $\boldsymbol{\Theta}$, respectively.[3] The gVAR model is estimated and compared based on the Extended Bayesian Information Criterion (EBIC; Chen & Chen, 2008). The sparsity of the solution is controlled by the EBIC hyperparameter γ, where higher values lead to more sparse networks. To increase the sensitivity for detecting nonzero edges in practical applications, γ is often set to 0 which reduces the EBIC to the regular BIC (Mansueto et al., 2020). Here, we set γ to either 0 or 0.5. We estimated regularized graphicalVAR models using the default grid of 50×50 penalty parameters for the temporal and contemporaneous coefficients. We additionally estimated unregularized graphicalVAR models by setting the LASSO shrinkage parameter 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. Specifically, we set the standard deviation of the prior distribution on the coefficients of the temporal network to $s_\beta \in \{0.2, 0.5, 1.0\}$ and those of the contemporaneous network to $s_\rho \in \{0.1, 0.3, 0.5\}$. We then created a grid of all possible prior combinations, resulting in nine different prior configurations. 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 $s_\rho = 0.3$). All other results are available in the supplement.

As the Bayesian estimation method does not perform structure estimation (i.e., it does not set coefficients to zero), 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). We used 90%, 95%, and 99% credible intervals and fitted the model using a Gibbs sampler with 50 burn-in and 50,000 sampling

[3] In the notation of Epskamp, van Borkulo, et al. (2018), our $\boldsymbol{\Theta}$ corresponds to their $\boldsymbol{K}$.

iterations which ensures convergence of MCMC samples. The large number of iterations was chosen because computational costs of the sampler were negligible and because the autocorrelation across iterations for the contemporaneous network was relatively high. For example, drawing 50,000 MCMC samples for a model with six nodes, 200 timepoints, and default priors required only about four seconds on an Intel i7-1260P processor. Convergence was checked visually via autocorrelations, trace plots, and effective sample sizes.

Performance Metrics. We based our choice of performance metrics on previous simulation studies investigating the performance of *graphicalVAR* (Mansueto et al., 2020). We calculated bias and correlation with the true parameters for all estimation methods. Bias refers to the mean absolute difference between the estimated and the true network. For the estimation methods that set coefficients to zero either by thresholding or by LASSO, we further investigated sensitivity (true-positive rate) and specificity (true-negative rate). For the *BGGM* method, we also computed the coverage rate of credible intervals, that is, the proportion of credible intervals containing the true value. To assess the precision of parameter estimates, we computed the width of credible intervals.

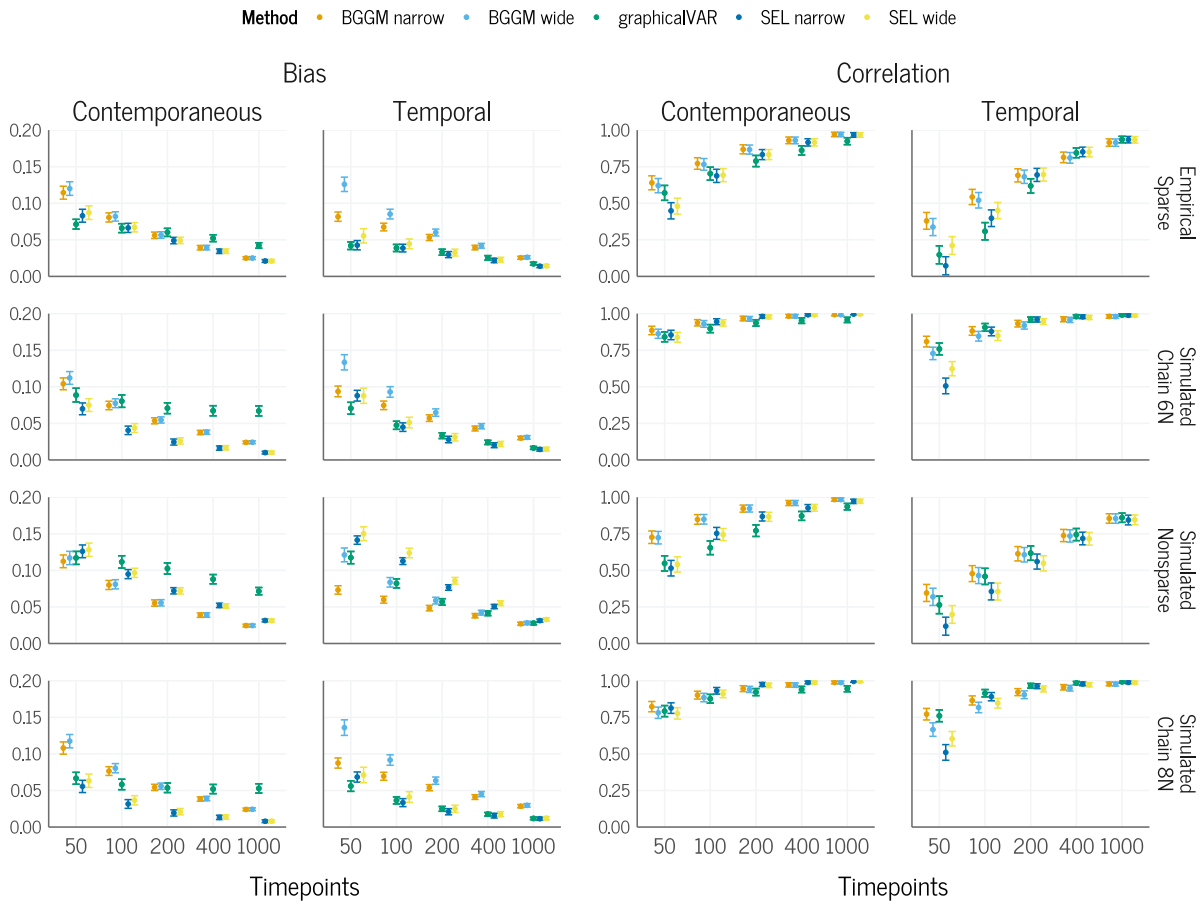
Results

Figure 2 shows the performance of Bayesian estimation for different priors with (denoted as 'SEL') and without (denoted as 'BGGM') thresholding (based on 95%-CIs). Moreover, the plot shows the results for LASSO estimation with the EBIC hyperparameter γ set to 0. Overall, the bias of the different methods was comparable in many conditions, with the largest differences occurring for small sample sizes. As expected, all estimation methods performed better and showed increasingly similar performance with more time points per individual. However, LASSO estimation showed a relatively smaller performance gain from more time points compared to other methods. At a small sample size of 50 time points, the overall performance of all methods was mediocre at best, with correlations falling below .50 for the temporal networks of the empirical sparse and simulated nonsparse graphs.

Which of the methods performed best depended on the data-generating process and

Figure 2

Bias and Correlation with True Edges for Different Simulation conditions.



Note. Separated by contemporaneous and temporal network (columns) and the data-generating processes (rows). 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. Vertical bars indicate $1.96 \times SE$.

the performance measure. In terms of bias, LASSO and thresholded Bayesian estimates with narrower priors performed best and showed very similar performance for sparse graphs, while non-thresholded Bayesian gVAR performed worse in most conditions with smaller sample sizes. In sparse graphs, there was no clear winner between thresholded Bayesian estimates and LASSO in terms of bias. However, in the non-sparse graph, non-thresholded Bayesian estimation with narrower priors performed best with regards to both bias and correlation. As the comparison of non-thresholded Bayesian estimation and LASSO regularization may be considered unfair when the data-generating process is dense, we also show results of

comparing the former method with nonregularized *graphicalVAR* in the supplement. The results show that Bayesian gVAR performed as well or better for all data-generating processes, especially when inducing stronger regularization via a slightly narrower prior.

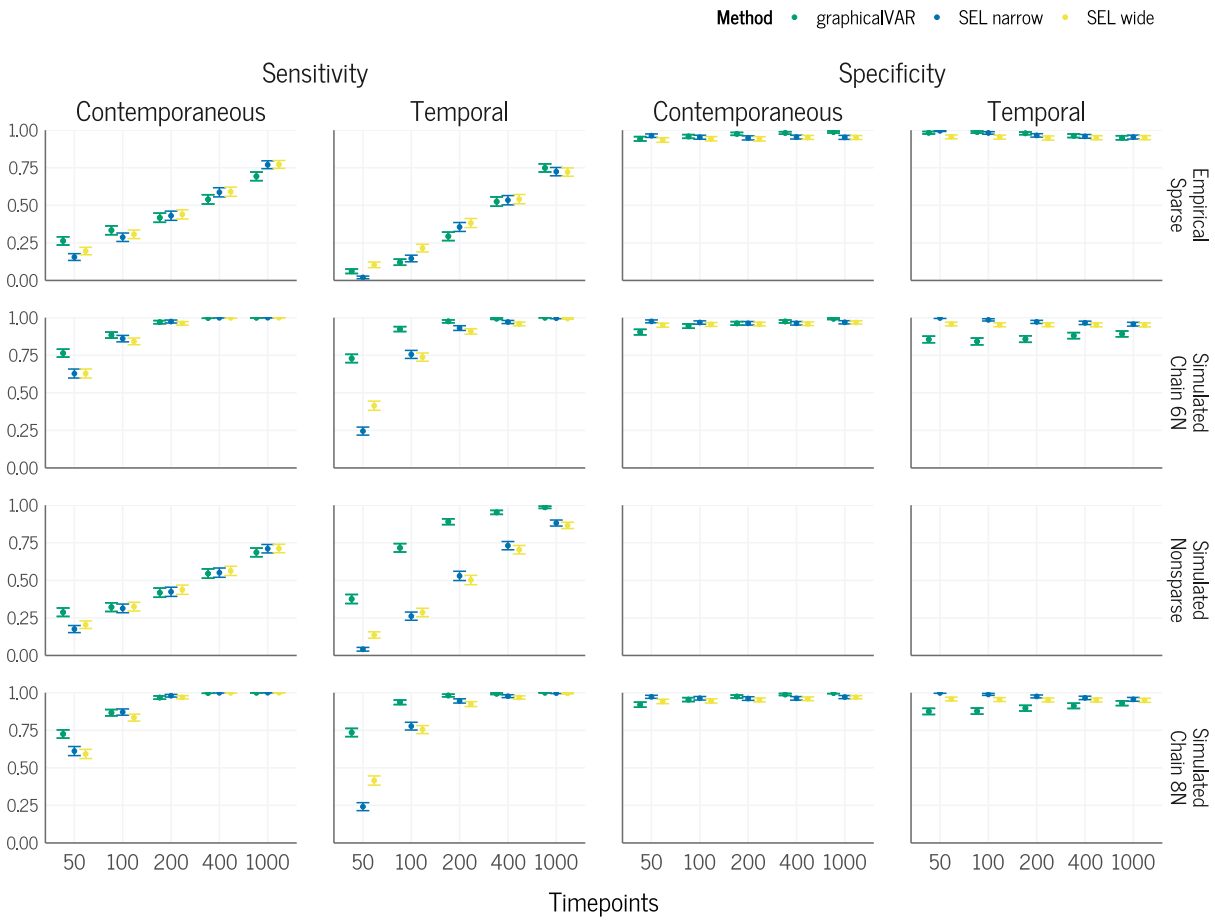
All methods performed better for contemporaneous networks than for temporal networks in terms of correlation, but not necessarily for bias, which is in line with previous simulations on the performance of LASSO gVAR (Mansueto et al., 2020). The difference was particularly striking in the nonsparse graph, where correlations were about twice as high in the contemporaneous as in the temporal network. However, there are more coefficients in the temporal than in the contemporaneous network, and thus, estimation may generally be more difficult.

For correlations, the thresholded Bayesian methods generally performed worse than other methods, especially for smaller sample sizes. With few time points, the non-thresholded Bayesian gVAR had a higher correlation with true parameters than other methods, probably due to the high degree of sparsity of both LASSO and thresholding when the sample size is small. The narrower and thus more informative prior performed as well as or even better than the more diffuse prior for both the non-thresholded and thresholded methods with regard to all performance criteria. When investigating other prior choices for our data, we generally found that a narrower prior on β and a relatively wider prior on ρ seemed to perform best. These additional results are presented only in the supplementary materials since we chose the specific hyperparameters for the 'wide' and the 'narrow' prior before analyzing the simulation results. Thereby, we avoid cherry-picking of an optimal prior setup for *BGGM* after seeing the results.

Figure 3 shows the specificity and sensitivity of LASSO estimation and Bayesian estimation with thresholding based on 95% CIs. Overall, Bayesian thresholding was conservative, leading to poor sensitivity below .50 in smaller sample sizes, especially in the temporal network. A sensitivity below .50 means that less than half of the true edges were detected by Bayesian thresholding, which partly explains its poor performance in terms of

Figure 3

Sensitivity and Specificity for Bayesian Thresholding and LASSO estimation.



Note. Estimation methods are shown in different colors in the same order as they appear in th legend. 'SEL' denotes Bayesian estimation with thresholding. Specificity for *Simulated Nonsparse* not shown, as there are no true-zero edges. Vertical bars indicate 95% CIs.

the correlation with true parameters reported above. For the temporal networks, LASSO estimation generally performed better. Regarding the contemporaneous networks, LASSO estimation also had a slightly better sensitivity for sample sizes up to 100, but became similar to Bayesian estimation for larger sample sizes. Conversely, Bayesian thresholding had a slightly higher specificity for most network structures in sample sizes up to 100. This was particularly pronounced in the temporal network of the chain graphs.

Although not necessarily a core objective of Bayesian inference, we further explored the coverage of credible intervals. Overall, coverage was good and close to the nominal values

of 90%, 95%, and 99%, with a median absolute difference of 0.6% between empirical coverage and credible interval width across all simulation conditions. With more time points, coverage decreased slightly for the temporal network and increased slightly for the contemporaneous network. The choice of prior distribution affected coverage only to a small degree while the effect was larger for smaller credible intervals. The supplementary materials contains plots of coverage and the width of credible intervals, as well as an assessment of prior sensitivity and different settings for *graphicalVAR*.

Summary

Overall, 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, continuous regularization 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.

In all conditions, a narrower, but not overly informative prior worked better than a wider, diffuse prior. Returning to the illustration of different priors in Figure 1, this result may not be surprising, as the wide prior places a considerable amount of prior probability on relatively large values that are rather implausible in temporal networks with standardized coefficients. We therefore recommend that researchers first examine the prior distribution for different sets of hyperparameters (similar as in Figure 1) in order to choose a suitable, possibly narrower prior than the default implemented in *BGGM*.

With respect to the frequentist properties of CIs, Bayesian estimates showed good coverage, suggesting that the posterior distribution provides a good indicator of estimation uncertainty. However, a major drawback of Bayesian estimation of gVAR models seems to be the potentially low sensitivity to detect non-zero edges in small sample sizes when using thresholding based on credible intervals. A 95% credible interval was often too wide, and hence, many edges that are nonzero in the population were set to zero. This may be due to the fact that the thresholding approach was not specifically designed for the task of reliable edge detection or strucutre selection (Sekulovski et al., 2023). Instead, thresholding based on credible intervals is a pragmatic ad-hoc solution as 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 that illustrates this trade-off can be found in the supplementary material. In general, our simulation provides limited information about sensitivity and specificity under various conditions, since we only focused on a limited number of hyperparameters for both Bayesian and LASSO gVAR.

We have shown that Bayesian estimation provides a viable alternative for estimating idiographic networks. Besides providing good estimation performance, Bayesian inference has the advantage of providing samples from the posterior distribution for all parameters. In part two of our manuscript, we develop a test that uses these samples for assessing heterogeneity and uncertainty of idiographic networks.

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., 2020). 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 two gVAR models, we do not set any edges to zero in order 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 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 $\boldsymbol{\Theta}$ the precision matrix of the contemporaneous network. In a first step, we compute two matrices that contain the differences of all parameters:

$$\boldsymbol{D}_B = \boldsymbol{B}_a - \boldsymbol{B}_b \tag{7}$$

$$\boldsymbol{D}_\Theta = \boldsymbol{\Theta}_a - \boldsymbol{\Theta}_b. \tag{8}$$

To quantify the amount of 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

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

in many applications such as LASSO (Tibshirani, 1996), network comparison methods (Tantardini et al., 2019), or in change detection for time series (Cabrieto et al., 2018). For our purpose, we chose three norms which 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 a value $\hat{d}$ that describes the estimated discrepancy between the networks for data sets A and B. 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 distribution 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 based on a single sample from the posterior. 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. 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 power 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 (see Discussion).

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

1. Estimate separate Bayesian gVAR models for the data sets a and b to obtain $s = 1, \ldots, S$ 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 $r = 1, \ldots, R$ 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 sampled from the posterior: $d_a^{(r)} = \|\boldsymbol{D}_a^{(r)}\|$ and $d_b^{(r)} = \|\boldsymbol{D}_b^{(r)}\|$

5. Compute posterior-based p-values as the proportion of posterior norms 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 $\boldsymbol{\Theta}$ instead of $\boldsymbol{B}$ in all steps.[5] We implemented the test in the `tsnet` package in R (https://github.com/bsiepe/tsnet) along with further functionality such as 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.

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,

[5] 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.

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 $\boldsymbol{\Theta}$ matrix by a factor of $\in \{1.4, 1.6\}$.[6] 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 $\boldsymbol{\Theta}$. Whether the value was subtracted or added was decided randomly until the resulting matrix fulfilled certain criteria relevant for 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 $\boldsymbol{\Theta}$ were 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.

Performance Metrics. In each simulation condition, we computed all possible pairwise network comparisons for all data sets. We only used 100 repetitions per condition, as 100 data sets already imply $\frac{100!}{(2!(100-2)!} = 4,950$ pairwise 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.

[6] 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.

Results

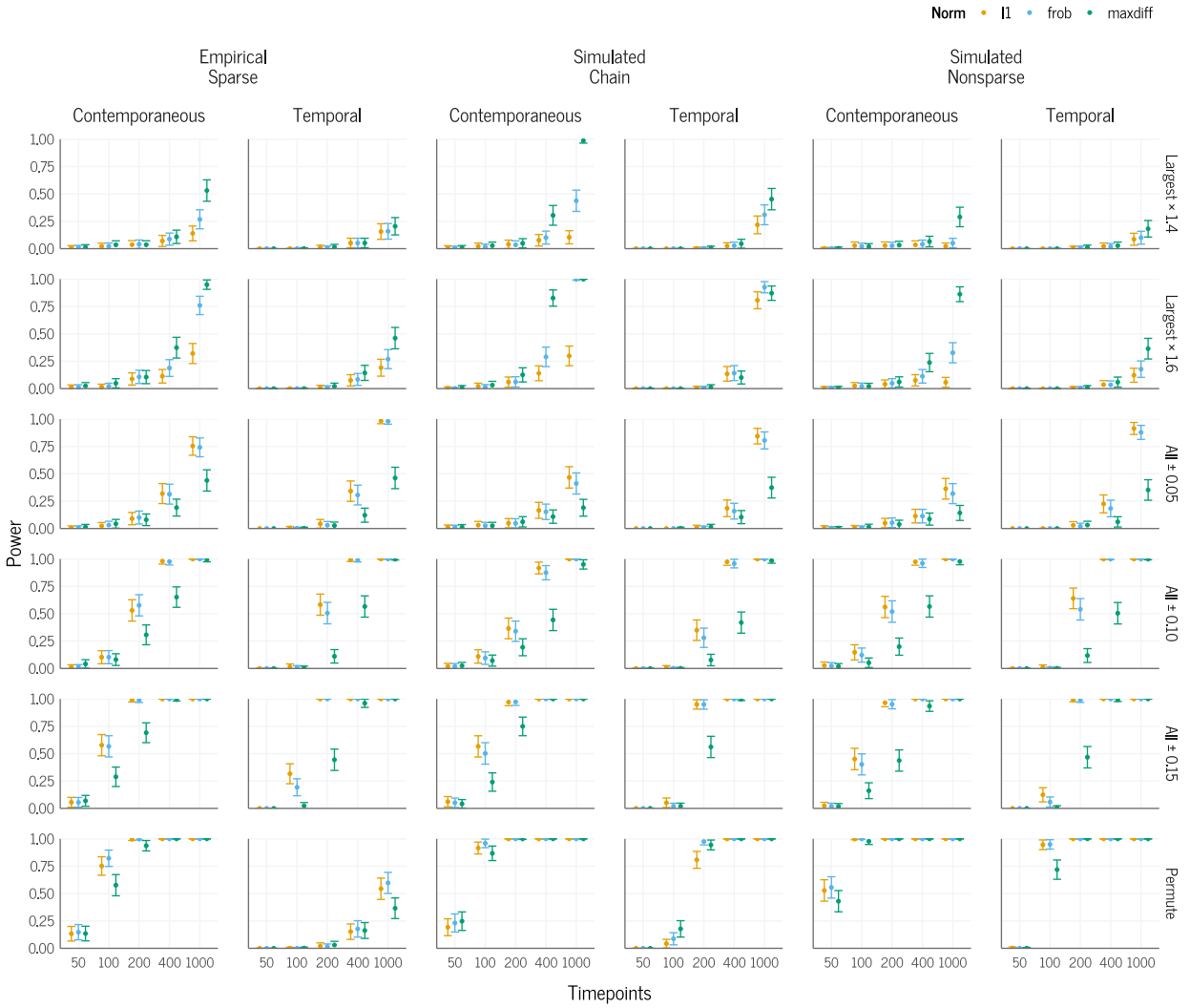
Figure 4 shows the power of the test (y-axis) with different numbers of time points (x-axis) across all manipulation conditions (rows) and data-generating processes (columns). Comparing the different rows shows that, the greater the difference between data-generating processes, the higher the power of 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 4 does not show a clear pattern regarding the performance of the test for different manipulations of the data-generating process. Overall, the test had a higher power to detect differences in the temporal network compared to the contemporaneous network for most change manipulations. This may be due to the fact that the temporal network consists of 36 parameters for six variables. In contrast, the contemporaneous network contains only 12 unique elements and thus provides less information for detecting differences. 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 5 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 4

Power of the Comparison Test for Idiographic Networks.

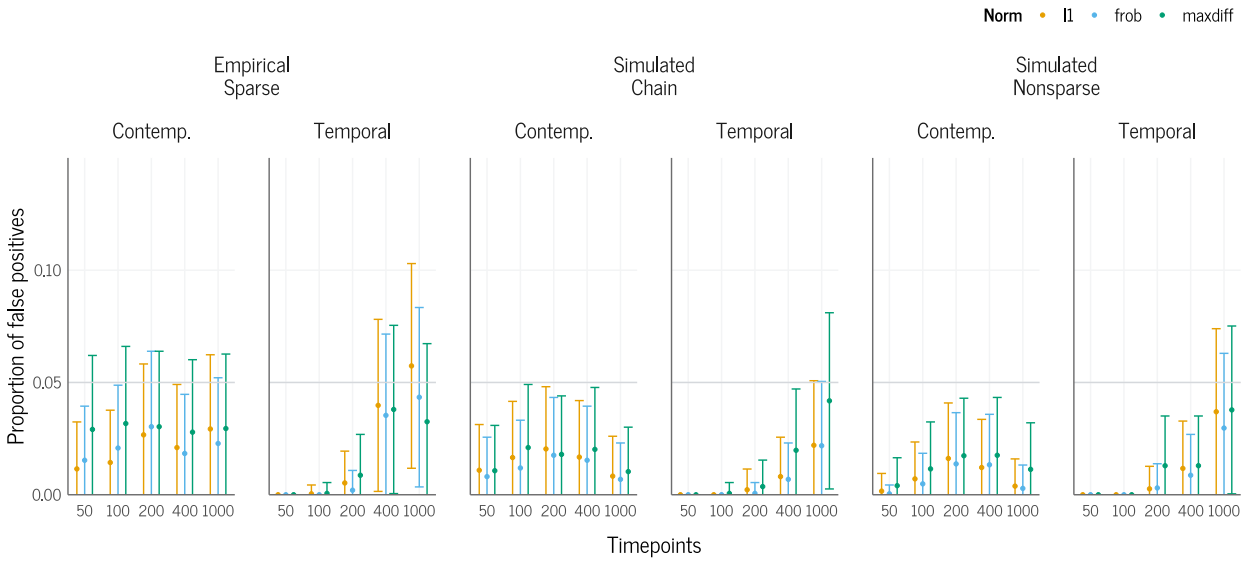


Note. Rows show different manipulations of the true, data-generating network. Power of the test to detect differences in idiographic networks with a prior of $s_\rho = 0.3$ and $\sigma_\beta = 0.2$. Vertical bars indicate $1.96 \times SE$, not adjusted for boundedness of the scale. Norms from left to right: ℓ_1-norms in orange, Frobenius norm in light blue, and maximum norm in green.

Prior Sensitivity

Figures 6 and 7 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. A narrower prior led to more conservative results (i.e., less evidence for the presence of differences) especially for smaller sample sizes. An

Figure 5
False-Positive Rate of the Test.



Note. Vertical bars indicate $1.96 \times SE$, 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.

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.

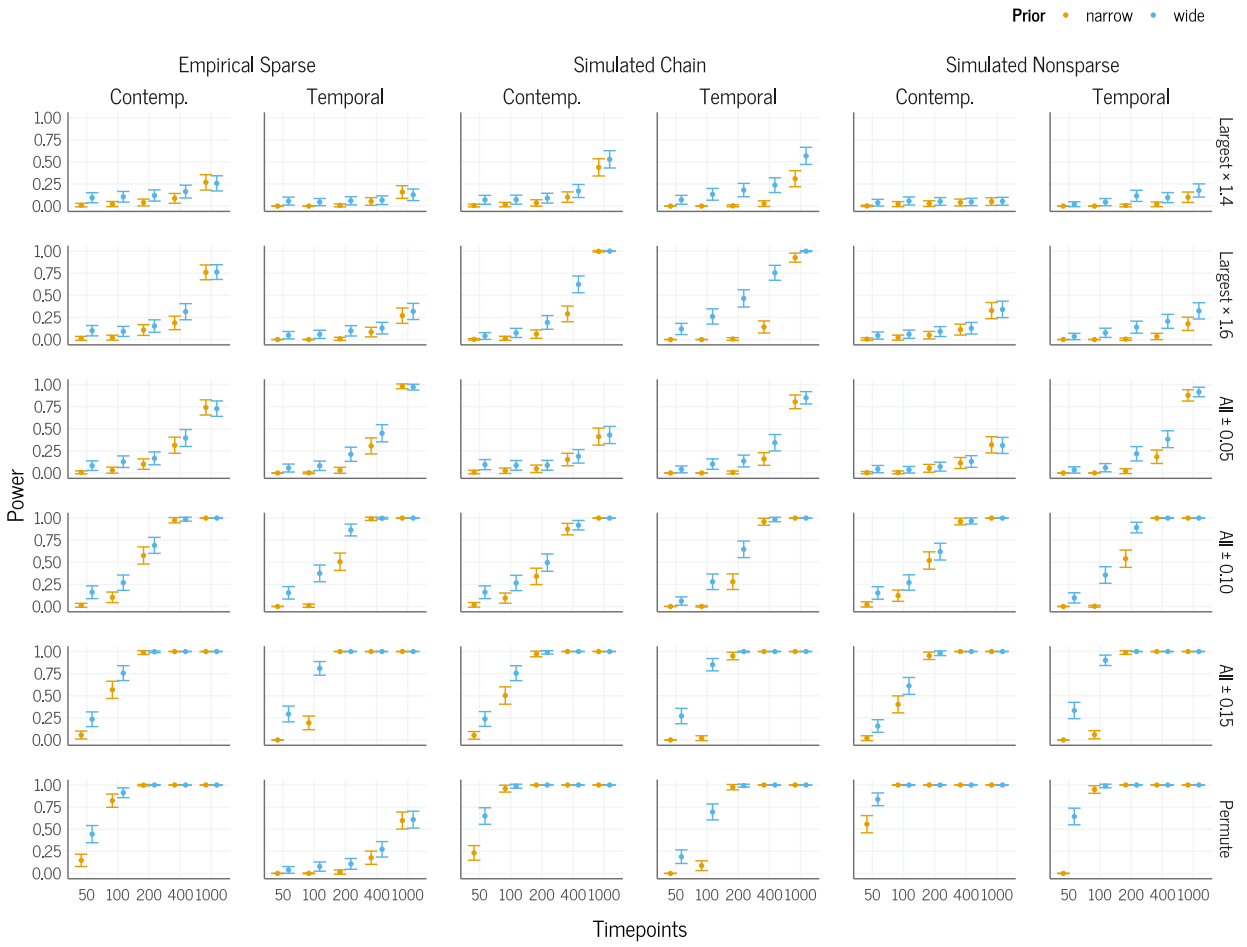
Summary

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. Whether differences in a few, specific edges or in all edges are plausible depends on the type of variables included in

Figure 6

Prior Sensitivity of the Power of the Test.

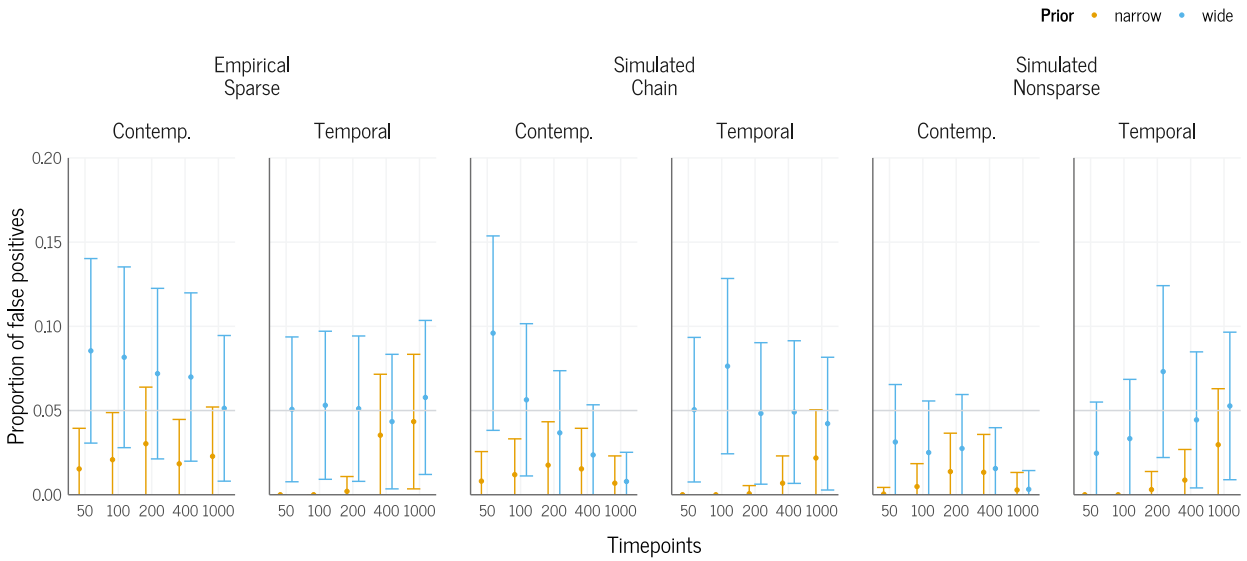


Note. Power of the comparison test in idiographic networks using only the Frobenius Norm. Narrow prior left (orange), wide prior right (light blue). Vertical bars indicate $1.96 \times SE$, not adjusted for boundedness of the scale.

the model and on the research question. 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 95%. 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, seems more plausible than differences in just one edge, we recommend using the Frobenius norm.

Figure 7

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 $1.96 \times SE$, not adjusted for boundedness of the scale.

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 realistic 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 prior to receiving psychotherapy. In these surveys, participants rated their 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 night time. This approach is further described and evaluated in a small simulation in Fisher et al. (2017).

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 95% while also exploring other norms for sensitivity analyses. Convergence diagnostics of MCMC sampling are implemented in *BGGM* and *tsnet* and shown in the supplement.

Results for Two Individuals

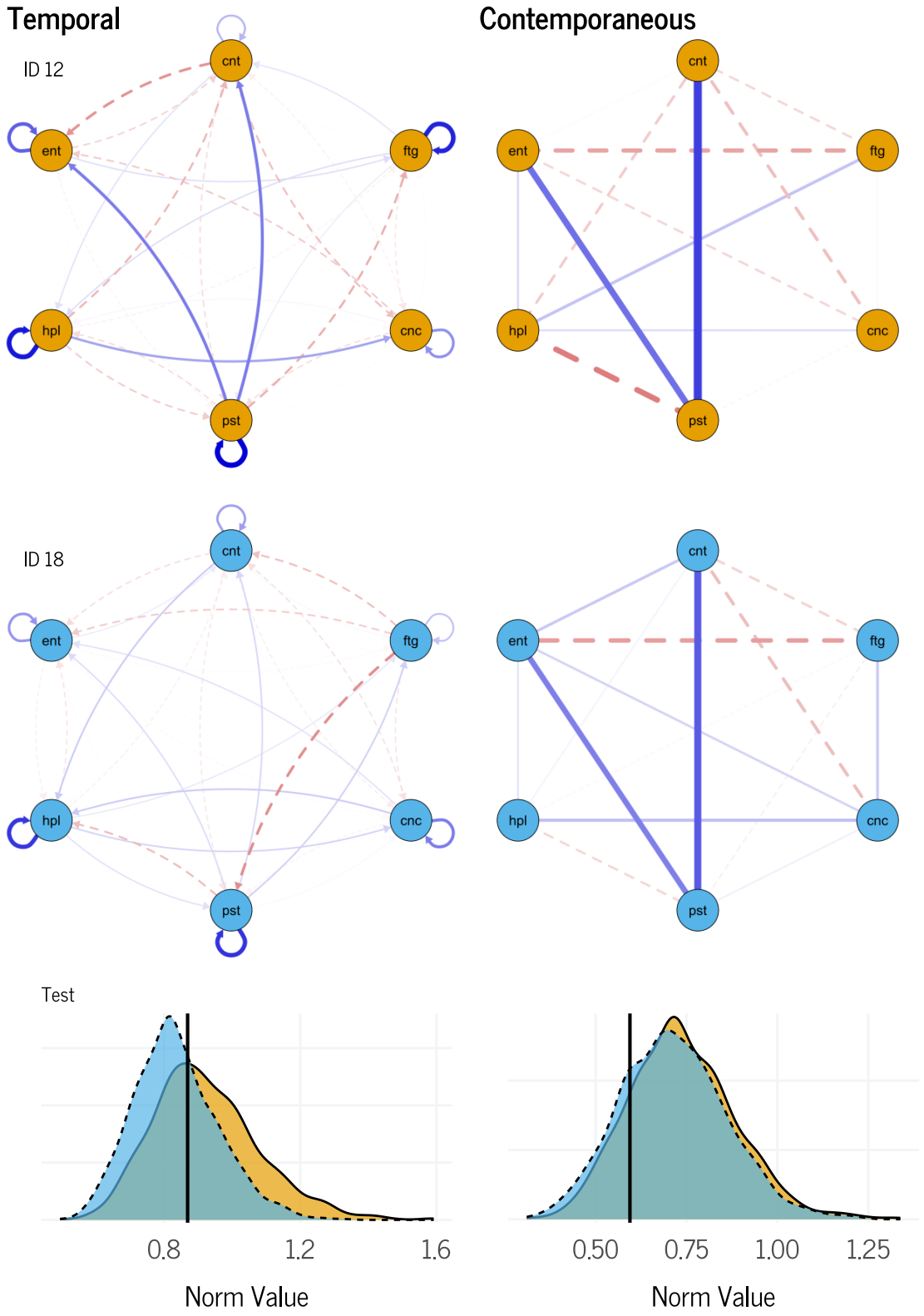
As an illustration, we present the results for two individuals for whom the comparison test did not indicate differences in either the temporal or the contemporaneous network. Figure 8 shows the estimated networks for participants 19 and 40 and a visualization of the comparison test. A visual inspection 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, regardless of whether the posterior distribution of participant 19 (in orange, solid line) or participant 40 (in blue, dashed line) served as reference. 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 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

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.

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 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 introduced Bayesian estimation 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. To set coefficients to zero with Bayesian gVAR, we used thresholding based on credible intervals (CIs). This approach performed well in terms of bias in several conditions. However, the Bayesian approach also lacked sensitivity in smaller sample sizes, especially for the temporal network, which is a relative advantage of LASSO gVAR. 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 specific edges (e.g., Marsman et al., 2022; Williams & Mulder, 2020).

Choosing an appropriate method for estimating gVAR networks 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 is a good method for 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 selected, 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 may outperform LASSO for dense networks.

Comparing Idiographic Network Models

In Part 2, we introduced a novel test to assess detect for 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 (or Euclidian) norm for network comparisons, 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 test and other useful functionality in the *tsnet* package in R.

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 is 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, major life event, or any other event of interest. More generally, if long time series are available, they may be split in half to estimate separate networks for the first and second segment. The new comparison test may be applied to test whether the two resulting networks are different.

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., 2020). Performance with sample sizes below 100 is generally weak regardless of the chosen method. 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 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). Such approaches which pool information across individuals likely perform better in many cases. Instead of replacing these methods, our test complements the methodological toolbox for network modeling.

Limitations and Future Research

Modeling idiographic dynamics of associations between psychological variables over time is a difficult task. Statistical models can at best provide an approximation of the underlying complex reality. This fact is reflected in common criticisms of (discrete-time) VAR models, most of which also apply to the gVAR model. These 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 (see Schuurman et al., 2015). We will not repeat these criticisms here as these have been discussed elsewhere (Bringmann, 2021; Haslbeck & Ryan, 2021; Ryan & Hamaker, 2021). However, these issues are also relevant for our empirical example. For instance, we accounted for non-equidistant

measurement intervals by cubic-spline interpolation (Fisher et al., 2017), also because the *BGGM* package is currently unable to handle overnight effects and ignores missing data. Extending the software to handle such issues is an important task for future work. Further, empirical data often contain non-normally distributed responses and different sample sizes between individuals, neither of which we have accounted for in our simulation studies.

The present work provides a first evaluation of the performance of Bayesian estimation for idiographic networks. Since the approach assumes a certain type of prior distributions, the range of available model specifications is limited. Other priors can be implemented in and are available for cross-sectional network estimation and time-series analyses in general. For example, spike-and-slab priors allow coefficients to be set to zero in a principled way (Marsman et al., 2022). Double exponential priors can be used to construct a Bayesian LASSO, which provides some advantages over regular LASSO in cross-sectional networks (Jongerling et al., 2022). Moreover, the econometric literature provides methods for Bayesian estimation and prior specification of VAR models (e.g., Giannone et al., 2015) and gVAR models (e.g., Ahelegbey et al., 2016; Paci & Consonni, 2020).

While our Bayesian estimation method accounted for parameter uncertainty, we did not account for overall structure uncertainty of the network. This is a potential extension of Bayesian estimation to be investigated in future work. In LASSO gVAR estimation, uncertainty about the network structure can be investigated using a block bootstrapping scheme described in Epskamp (2020). In a Bayesian framework, Bayes factors could be developed to evaluate the evidence for and against the inclusion of a certain edge. Bayesian model averaging could then 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, the *BGGM*-package allows researchers to implement

edge-specific informative priors (Jongerling et al., 2022). As noted above, the estimation of gVAR models still suffers from limited power 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, Li et al., 2022) or (Hamaker et al., 2018)), these could be improved with further work on prior choices and structure uncertainty. Also, as others have already pointed out, highly data-driven methods such as the aforementioned GIMME could benefit from Bayesian ways of quantifying uncertainty (Nestler & Humberg, 2021). Further, centrality values are often of major interest for interpreting the results of network models in clinical settings (Bringmann, 2021). Future research could explore the performance of uncertainty estimates for centrality measures, as Bayesian methods have proved to be promising for this task in cross-sectional networks

A main conceptual limitation of the proposed comparison test concerns its inability to provide evidence for the null hypothesis. Negative results of the test can occur both due to large estimation uncertainty or due to the actual invariance of data-generating processes. To separate these two causes, one may develop Bayes factors for testing differences between networks. Comparison tests for idiographic networks may also be constructed in different ways as briefly sketched in the following. First, one may rely on posterior predictive tests (Williams et al., 2020). Second, 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. In addition, classic, frequentist invariance tests could be used when fitting gVAR models in a SEM framework (e.g., Fisher et al., 2017) or in the *psychonetrics*-package (Epskamp, Borsboom, et al., 2018). Finally, simulation-based procedures similar to ours could be performed via parametric bootstrapping for frequentist models. 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 Bayesian inference 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 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.

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 Bandiera_abtest: a Cg_type: Nature Research Journals Primary_atype: Reviews Subject_term: Scientific data;Statistics Subject_term_id: scientific-data;statistics

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

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771. https://doi.org/10.1093/biomet/asn034

Costantini, G., Kappelmann, N., & Epskamp, S. (2021). EstimateGroupNetwork: Perform the Joint Graphical Lasso and Selects Tuning Parameters.

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). graphicalVAR: Graphical VAR for Experience Sampling Data.

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

Fried, E. I., Papanikolaou, F., & Epskamp, S. (2021). Mental health and social contact during the covid-19 pandemic: An ecological momentary assessment study. *Clinical Psychological Science*, *10*(2), 340–354. https://doi.org/10.1177/21677026211017839

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

Hall, M., Wagner, A. A., Scherner, P., Michael, K. L., Lawyer, G., Lutz, W., & Rubel, J. (2022). Using Personalized Assessment and Network Model Feedback in Psychotherapy: Proof of Principle for the TheraNet Project. https://doi.org/10.31234/osf.io/8deyj

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). 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.

Hoekstra, R. H. A., Epskamp, S., & Borsboom, D. (2022). Heterogeneity in individual network analysis: Reality or illusion? *Multivariate Behavioral Research*, *0*(0), 1–25. https://doi.org/10.1080/00273171.2022.2128020

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

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

Levinson, C. A., Hunt, R. A., Keshishian, A. C., Brown, M. L., Vanzhula, I., Christian, C., Brosof, L. C., & Williams, B. M. (2021). Using individual networks to identify treatment targets for eating disorder treatment: A proof-of-concept study and initial data. *Journal of Eating Disorders*, *9*(1). https://doi.org/10.1186/s40337-021-00504-7

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., van Weert, J. C. M., Schouten, B. C., & Epskamp, S. (2020). Investigating the feasibility of idiographic network models. https://doi.org/10.31234/osf.io/hgcz6

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. (2023). 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

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mulder, J., & Pericchi, L. R. (2018). The Matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis*, *13*(4), 1193–1214. https://doi.org/10.1214/17-BA1092

Nestler, S., & Humberg, S. (2021). Gimme's ability to recover group-level path coefficients and individual-level path coefficients. *Methodology*, *17*(1), 58–91. https://doi.org/10.5964/meth.2863

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

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). 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

Rothman, A. J., Levina, E., & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, *19*(4), 947–962.

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

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*(2-3), 185–206. https://doi.org/10.1080/00273171.2015.1065398

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., van Bork, R., van den Bergh, D., & Marsman, M. (2023). Testing conditional independence in psychometric networks: An analysis of three bayesian methods. https://doi.org/10.31234/osf.io/ch7a2

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

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 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 de Schoot, R., Veen, D., Smeets, L., Winter, S. D., & Depaoli, S. (2020). A tutorial on using the wambs checklist to avoid the misuse of bayesian statistics. In *Small Sample Size Solutions*. Routledge.

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). BGGM: Bayesian Gaussian Graphical Models.

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 A

MCMC Sampling

Let $\boldsymbol{Y}$ be the matrix of observed data with n rows for k dependent variables and $\boldsymbol{X}$ be the matrix of p lag-1 predictor variables (i.e. $\boldsymbol{Y}$ shifted back by one time point). Again, as stated in the manuscript, we deviate from the notation of Mulder and Pericchi (2018) by using $\boldsymbol{C}$ instead of $\boldsymbol{B}$ for the scale matrix of the matrix-F prior. Gibbs sampling is then performed according to the following steps:

1. Obtain the scatter matrix of independent variables as the sum of the outer product of the independent variables and the prior matrix,

$$\boldsymbol{S_X} = \boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{\beta_{prior}},$$

where

$$\boldsymbol{\beta_{prior}} = \frac{1}{(s_\beta)^2} \cdot \boldsymbol{I_k}.$$

2. Initialize MCMC sampling by setting the starting value of $\boldsymbol{\Sigma}$ to the sample correlation matrix of the observed data and the starting value of $\boldsymbol{\Theta}$ to the inverse of the sample correlation matrix of the observed data (i.e. the sample precision matrix). $\boldsymbol{\Psi}$ is set to the identity matrix.

3. Draw a matrix with β-coefficients of the temporal network from a multivariate normal distribution,

$$\boldsymbol{B} \sim \mathcal{MVN}(\boldsymbol{S_X^{-1}} \boldsymbol{X}^\intercal \boldsymbol{Y}, \boldsymbol{\Sigma} \otimes \boldsymbol{S_X^{-1}})$$

4. Compute the scatter matrix of dependent variables

$$\boldsymbol{S_Y} = \boldsymbol{Y}^\intercal \boldsymbol{Y} + \boldsymbol{I_k} - \boldsymbol{B}^\intercal \boldsymbol{S_X} \boldsymbol{B}$$

5. Draw a new sample for the matrix $\boldsymbol{\Psi}$,

$$\boldsymbol{\Psi} \mid \boldsymbol{\Theta} \sim \mathcal{W}\left((\boldsymbol{C} + \boldsymbol{\Theta})^{-1}, \nu_{MP} + \delta_{MP} + k - 1\right)$$

where

$$\boldsymbol{C} = \epsilon \cdot \boldsymbol{I_k}$$

$$\epsilon = 0.001$$

$$\nu_{MP} = \delta + k - 1$$

$$\nu = \epsilon^{-1}$$

$$\delta_{MP} = \nu - k + 1$$

This is explained further in Williams and Mulder (2020) and Williams et al. (2020).

6. Draw a new sample for the precision matrix $\boldsymbol{\Theta}$,

$$\boldsymbol{\Theta} \mid \boldsymbol{\Psi} \sim \mathcal{W}((\boldsymbol{\Psi} + \boldsymbol{S_Y})^{-1}, (\nu + (n - 1))$$

7. Compute the covariance matrix $\boldsymbol{\Sigma}$ corresponding to the contemporaneous network,

$$\boldsymbol{\Sigma} = \boldsymbol{\Theta^{-1}}.$$

Appendix B

Data-Generating Processes

The following plots display the data-generating matrices we used to genenerate data. Temporal directed networks are on the left, contemporaneous undirected networks 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.

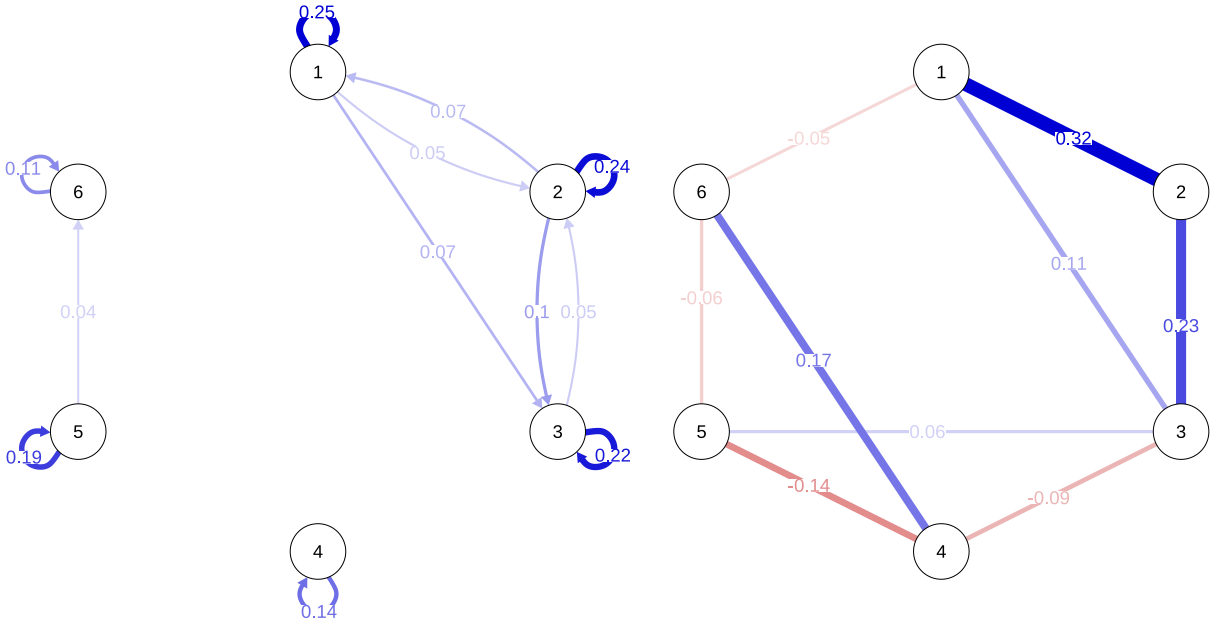


Figure B1
Empirical Sparse Network

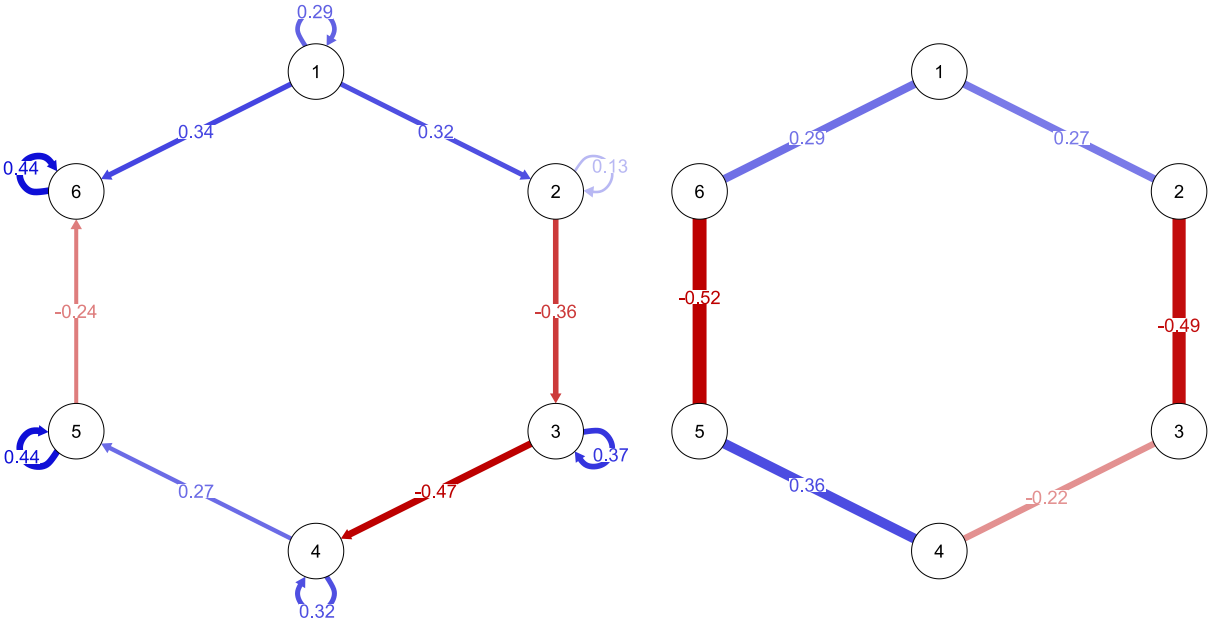


Figure B2

Sparse Chain with 6 Nodes

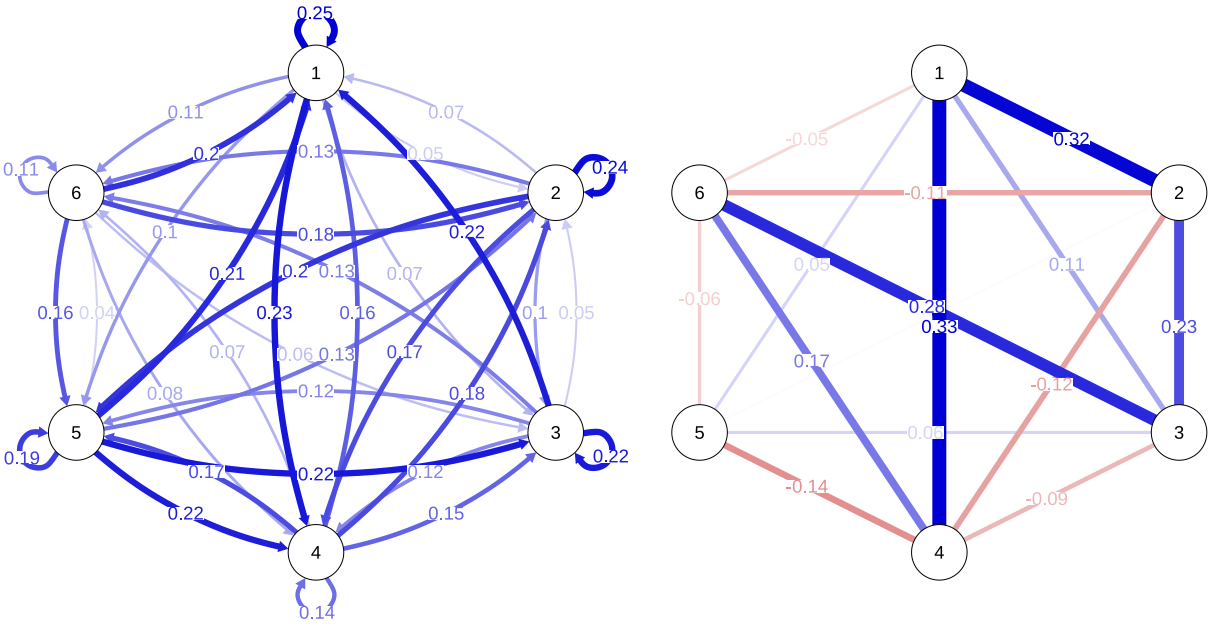


Figure B3

Nonsparse Network

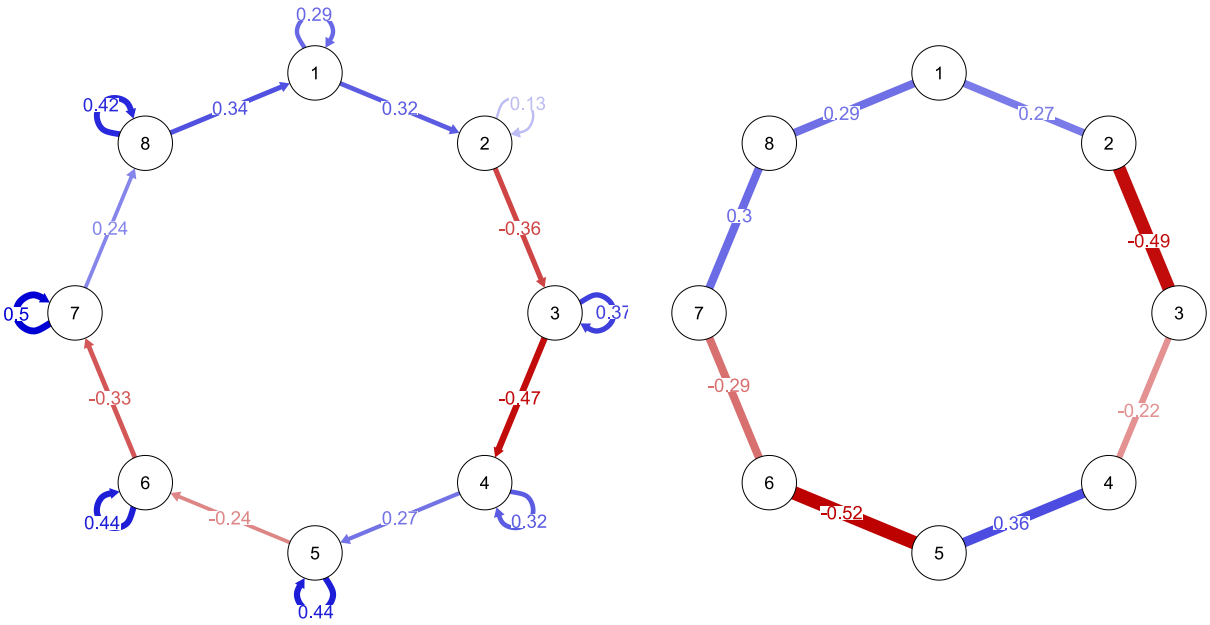


Figure B4

Sparse Chain with 9 Nodes