UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

# Testing for group differences in multilevel vector autoregressive models

Haslbeck, J.M.B.; Epskamp, S.; Waldorp, L.J.

DOI
10.31234/osf.io/dhp8s
10.3758/s13428-024-02541-x

Publication date
2025

Document Version
Final published version

Published in
Behavior Research Methods

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
Haslbeck, J. M. B., Epskamp, S., & Waldorp, L. J. (2025). Testing for group differences in multilevel vector autoregressive models. *Behavior Research Methods*, *57*(3), Article 100. https://doi.org/10.31234/osf.io/dhp8s, https://doi.org/10.3758/s13428-024-02541-x

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.
*UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)*

Download date:14 Nov 2026

Behavior Research Methods (2025) 57:100
https://doi.org/10.3758/s13428-024-02541-x

ORIGINAL MANUSCRIPT

# Testing for group differences in multilevel vector autoregressive models

**Jonas M. B. Haslbeck**[1,2] · **Sacha Epskamp**[3] · **Lourens J. Waldorp**[2]

Accepted: 4 September 2024 / Published online: 20 February 2025
© The Psychonomic Society, Inc. 2025

**Abstract**
Multilevel Vector Autoregressive (VAR) models have become a popular tool for analyzing time series data from multiple subjects. Many studies aim to investigate differences in multilevel VAR models between groups, such as patients and healthy controls. However, there is currently no easily applicable method to make inferences about such group differences. Here, we present two standard tests for making such inferences: a parametric test and a nonparametric permutation test. We explain the rationale for both tests, provide an implementation based on the popular R-package *mlVAR*, and evaluate their performance in recovering group differences in scenarios resembling empirical research using a simulation study. Finally, we provide a fully reproducible R-tutorial on testing for group differences in a dataset of emotion measures using the new R-package *mnet*.

**Keywords** Time series modeling · Vector Autoregressive models · VAR models · Network models · Group comparison

## Introduction

Time series are becoming a common data type across disciplines in psychological research (Kuppens, Dejonckheere, Kalokerinos, & Koval, 2022; Miller, 2012; Conner & Barrett, 2012; Trull & Ebner-Priemer, 2014; Hamaker & Wichers, 2017). A natural first approximation of the temporal dynamics is the Vector Autoregressive (VAR) model, which models each variable as a linear combination of all variables (including itself) at previous time points (Hamilton, 1994). However, in many situations, one does not have enough data from a single subject to estimate the VAR model (Mansueto, Wiers, van Weert, Schouten, & Epskamp, 2022; Dablander, Ryan, & Haslbeck, 2020; Bulteel, Mestdagh, Tuerlinckx, & Ceulemans, 2018), which is one of the reasons why many researchers are estimating multilevel VAR models (Epskamp, Waldorp, Mõttus, & Borsboom, 2018; McNeish & Hamaker, 2020).

A central question in many studies is how the dynamics between variables across time differs across groups of subjects. For example, Elovainio, Kuula, Halonen, and Pesonen (2020) investigated how multilevel VAR models fitted to self-reported emotion measures differed between individuals with and without sleep problems; Curtiss, Fulford, Hofmann, and Gershon (2019) compare multilevel VAR models fitted to affect and physical activity in individuals diagnosed with bipolar disorder and healthy controls; and Snippe et al. (2017) compared multilevel VAR models fitted to emotion measures before and after treatment for major depression.

However, in these and similar studies group differences are typically assessed by (visually) comparing the parameter estimates and no formal inference about group differences is performed. This has the disadvantage that we have no way to separate signal from noise, that is, to decide whether observed group differences occur in the population or whether they are due to sampling variation. This problem is an example of the more general issue of distinguishing population inter-individual heterogeneity from sampling variation (Hoekstra, Epskamp, & Borsboom, 2022).

In this paper, we address this issue by considering two tests for group differences in multilevel VAR models. We lay out the rationale for a parametric test and a nonparametric per-

✉ Jonas M. B. Haslbeck
jonashaslbeck@protonmail.com; www.jonashaslbeck.com

1 Department of Clinical Psychological Science, Maastricht University, Maastricht, The Netherlands

2 Department of Psychological Methods, University of Amsterdam, Nieuwe Achtergracht 129-B, Postbus 15906, 1018 WT Amsterdam, The Netherlands

3 Department of Psychology, National University of Singapore, Singapore, Singapore

mutation test and discuss their advantages and disadvantages based on theory. We then determine how well group differences can be recovered with both tests in a set of scenarios mirroring applied research settings using a simulation study. To make the tests available, we present the new R-package *mnet*, which implements both tests and is based on the R-package *mlVAR* (Epskamp et al., 2018), which is currently the most popular open source software package to estimate multilevel VAR models. Finally, we provide a fully reproducible tutorial in which we test differences in emotion dynamics across groups with high vs. low depressive symptoms.

## Parametric and nonparametric tests for group differences in multilevel VAR models

In this section we briefly discuss the multivariate Vector Autoregressive (VAR) model. Next, we discuss parametric and nonparametric tests for differences in VAR parameters across two groups.

### Multilevel Vector Autoregressive (VAR) models

In the standard lag-1 VAR model, each variable is predicted by itself and all other variables at the previous time point. Specifically,

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{B}\left(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\zeta}_t,$$

where $\boldsymbol{y}_t$ is a vector of $p$ variables at time point $t$, which is modeled by a $p$-vector of means $\boldsymbol{\mu}$, and a linear combination of the same (mean-centered) variables at the previous time point $\boldsymbol{B}\boldsymbol{y}_{t-1}$, where $\boldsymbol{B}$ is a $p \times p$ matrix of lagged effects, and $\boldsymbol{\zeta}_t$ is a $p$-vector of Gaussian innovation.

Since in many psychological studies the number of available measurement points is small relative to the number of parameters of the VAR model, estimating a VAR model from the data of a single individual is often unfeasible (Mansueto et al., 2022; Dablander et al., 2020; Bulteel et al., 2018). Many researchers therefore use multilevel VAR models, where parameters of each subject VAR model are shrunk towards the grand means of parameters across all subjects, which leads to estimates with less variance (McNeish & Hamaker, 2020; Epskamp et al., 2018; McElreath, 2018). This model takes the same form as the model above, except that we introduce person indices for means ($\boldsymbol{\mu}_s$), lagged effects ($\boldsymbol{B}_s$), and the distribution of residuals ($\boldsymbol{\zeta}_{t,s}$).

There are currently two popular approaches to estimating multilevel VAR models. One is the DSEM framework in M*plus* (Asparouhov, Hamaker, & Muthén, 2018), which fits the multilevel VAR model as a joint model.

The alternative is the R-package *mlVAR* (Epskamp et al., 2018), which breaks estimation up in several steps. The method begins by approximating population within-person means with within-person sample means. In a first step, the method then estimates the lagged parameters with $p$ separate regressions, in which $y_t i$ is modeled with univariate multilevel regression model with the set of within-person lagged variables $\boldsymbol{y}_t - 1$ and the personwise sample means $\bar{\boldsymbol{y}}$ as predictors. In a second step, the level 2 regression coefficients associated with personwise means are transformed into partial correlation to obtain an estimate for partial correlations between random intercepts (obtaining the "between-network"; for details see Epskamp et al., 2018). In a third step, residuals from the $p$ models from step one are predicted in another $p$ univariate multilevel regressions in which the residual of $y_{ti}$ is predicted using residuals from all other variables at the same time point. The resulting regression coefficients are again transformed into partial correlations. The estimation of partial correlations between random intercepts (step 2) and partial correlations between residuals (step 3) are a disjoint likelihood approach which leads to two estimates for a given parameter. These estimates are combined by taking their average. For more details on this method, see Supplement S1 in Epskamp et al. (2018).

While this approach is fast and available in open-source software, the fact that the fitted model is misspecified (using sample means as proxies for population means and using a disjoint pseudolikelihood approach in steps two and three) with respect to a joint multilevel VAR model has several disadvantages. For example, not all random effects can be correlated and parameters are subject to a number of biases, especially if the number if time points per person is low (McNeish & Hamaker, 2020; Haslbeck & Epskamp, 2023b; Hamaker, 2022). As we will see later, these biases will also have consequences for the quality of the test we develop on top of the *mlVAR*-package. Nonetheless, we use this package in our paper since it is the one that is openly available.

The parameters that are typically of main interest in multilevel VAR models are the fixed effects of the lagged coefficients and their random effects variances, the fixed effects of relations between innovations and their random effects variances (often called contemporaneous network), and the relations between subjects-specific means (often called between-subjects network). While the model also includes fixed effects and random effects for means and random effects correlations between lagged coefficients, these typical not of central interest and we therefore focus below on the five mentioned types of parameters discussed above.

### Parametric test

Making inferences about group differences in a Frequentist setting typically involves testing an observed group difference against the null hypothesis that there is no group difference in a parametric sampling distribution (e.g., the

normal distribution for the mean parameter). This is also the approach we take in this paper. To perform such a test, we need a sampling distribution for the group difference of a given parameter in the multilevel VAR model under the null hypothesis that both groups are identical. Based on the sampling distribution of the parameter (or parameter difference), we can compute confidence intervals and *p*-values and perform statistical inference on the group differences.

One way to obtain these sampling distributions is to use standard theory (Central Limit Theorem) and assume that sampling distributions are approximately Gaussian. In the context of group differences in parameters in the multilevel VAR model, this means that we can use the observed group difference and the standard errors of the parameters in the two groups to obtain a confidence interval and a *p* value for the group differences. In general, the population standard errors of estimates are unknown, which is why Student's *t* distribution should be preferred over the Gaussian distribution, especially for small sample sizes. In addition, we generally do not want to assume that the standard errors are equal across to groups, especially when the compared groups have different sizes. We therefore suggest to use Welch's *t* test, which takes unequal variances into account.

### Nonparametric permutation test

In a permutation test, we construct a sampling distribution under the null hypothesis that there are no group differences by randomly shuffling the subjects in the two groups, while keeping the original group sizes as in the empirical data. In the case where the groups are dependent, for example when measuring the same subjects twice before and after a treatment, the permutations are performed within individuals.

Repeating this process $N_P$ times results in $N_P$ pairs of datasets for which we have no group differences in expectation. We can therefore use them to construct an empirical sampling distribution. We estimate a multilevel VAR model from each dataset in a pair and compute the difference in each parameter across the two models. Taking the differences from all $N_P$ pairs together, this gives us a sampling distribution for each parameter given the null hypothesis that all group differences are equal to zero. We can then compare the group differences in the original groups against these sampling distributions to compute empirical *p* values and use them for inference. The approach taken here is similar to the one in the Network Comparison Test introduced by Van Borkulo et al. (2022) for testing group differences across cross-sectional network models (in a fixed effects analysis).

While we are here primarily interested in differences in the means of parameters across groups, a permutation test is technically a test for the equality of distributions, which here means the equality of the distribution of any given parameter across the two groups. To evaluate equality we summarize the distributions by their means, which requires the assumption that under the null hypothesis the distributions in both groups have the same shape or are both symmetric (e.g., Good, 2013). This is required to use the test based on the mean to draw conclusion about the equality of the distributions.

Performing a permutation test requires specifying the number of permutations. More permutations lead to more precise estimates of population sampling distributions, however, this increase in precision has to be balanced against the computational cost of additional permutations. While this computational cost is often negligible, it is considerable in the context of multilevel (VAR) models. In Appendix A we discuss how many permutations are necessary to obtain good estimates of sampling distributions and *p* values, drawing both from statistical theory on permutation tests and simulations in the context of VAR models. While any cutoff on the number of permutations is necessarily arbitrary, we conclude that $N_P = 1000$ permutations generally already lead to acceptable estimates of *p* values.

### Parametric vs. nonparametric permutation test

Parametric tests tend to have more power when all their assumptions are met, that is, the sampling distribution is indeed a Gaussian distribution and the data are on a continuous scale. However, these assumptions may often be violated in empirical research. While the Central Limit Theorem guarantees that many sampling distributions will converge to a Gaussian distribution when the sample size (here the number of subjects $N_{subj}$) is large enough (given certain assumptions, see e.g., Vaart, 1998), the sampling distribution in empirical data may considerably deviate from a Gaussian distribution if the distributions of person-specific parameters are non-Gaussian and if $N_{subj}$ is relatively low. In addition, many time series datasets use Likert scales with ordinal 3-7 response categories, which clearly do not lead to continuous measurements. In these cases, a permutation test may be more appropriate. Finally, a permutation test can also be used if no standard error can be obtained for a given test-statistic. In the case of the *mlVAR* package no standard errors are available for random effects variances and consequently the parametric test cannot be used to make inference about group differences in these parameters.

## Performance of tests in recovering group differences

We will use a simulation study to evaluate to what extent the parametric test and the nonparametric permutation test are able to recover group differences of different sizes in

multilevel VAR models in scenarios that are typical in applied research.

## Simulation setup

### Defining true models & data generation

We generate data from two multilevel VAR models that differ in the most relevant types of parameters and evaluate the extent to which we can recover these group differences as a function of the size of group differences $\Delta \in \{0.04, 0.08, 0.12\}$, the total number of subjects $N_{\text{subj}} \in \{50, 100, 200\}$ and the proportion of subjects in the first group $p_{\text{G1}} \in \{0.2, 0.5\}$. We chose these ranges because they cover the typical ranges found in empirical research. We fixed the number of variables to $p = 8$ since this is in the range of variables typically modeled in VAR models in psychological research. We also fixed the number of time points for each subject to $N_t = 100$. It would have also been desirable to vary parameters such as the number of variables, the number of time points $N_t$ for each subject, or the size of random effects variances. However, due to the considerable computational cost of the permutation test for the multilevel VAR model, we had to limit the simulation study to a small number of factors.

We construct the pair of data-generating multilevel VAR models in the following way: we first define a model for the first group; then we define group differences which relative to the first group define the second group. When specifying the VAR model of the first group, we mimic VAR models often found in Ecological Momentary Assessment (EMA) data. EMA time series often capture positive and negative affect items, which show the following stable patterns in the $p \times p$ matrix $\boldsymbol{B}$ specifying temporal effects with a lag of 1: autocorrelations are positive and much larger than cross-lagged effects, cross-lagged effects within valence are positive, and cross-lagged effects between valence are negative (for a review see Ryan, Dablander, & Haslbeck, 2023). We implemented this by setting all autocorrelations $\beta_{i,i} = 0.4$. We choose off-diagonal cross-lagged effects with absolute value 0.075. We chose those values to represent VAR models typically found in empirical data and to ensure that the true VAR models were stable (i.e., all eigenvalues of $\boldsymbol{B}$ are within the unit circle; see e.g., Hamilton, 1994). We set the grand mean of all variables to zero. The innovation variances of all variables for all subjects are set to 1. We choose partial correlations between innovations with absolute values 0.075. Again, we specify a block structure as above for the lagged effects, where partial correlations within valence are positive, and partial correlations between valence are negative.

Importantly, we choose to model heterogeneity in parameters across individuals within each group with a Gaussian distribution. We choose this scenario which favors the parametric test to see how much worse the nonparametric permutation test performs compared to it. If it performs equally well or comparably, this would lead us to recommend the permutation test, because if performs comparably when the assumptions of the parametric test are met, and will perform better when they are not met. For lagged effects and innovation partial correlations we introduce heterogeneity across subjects with a Gaussian distribution with standard deviation $\sigma_{\text{RE}} = 0.05$. For the intercepts, we introduce heterogeneity across subjects by drawing from a Gaussian distribution with partial correlations with the same block structure as the innovation partial correlations, and with standard deviations of $\sigma_{\text{RE}} = 0.50$, since random effects variances of intercepts tend to be much larger than random effects variances of association parameters in psychological time series data.

Next, we introduce group differences in (1) partial correlations between person-means (i.e., the "between-network"), (2) the fixed and (3) random effects variances of the innovation partial correlations, and (4) the fixed and (5) random effects of the lagged effects. For each parameter type, we randomly choose 15% parameters for which we introduce a group difference. Rounding to the nearest integer, this gives us a fixed number of five group differences for partial correlations between means and fixed and random effects variances of innovation partial correlations, and a fixed number of ten for group differences for fixed and random effects variances for lagged effects. For random effects standard deviations we add the group difference to the respective random effects standard deviations of Group 1 to avoid negative variances. For the remaining three types of parameters we randomly choose the sign of the group difference.

VAR matrices are required to have eigenvalues within the unit circle, and covariance matrices both for innovations and random intercepts are required to be positive definite. In case the above procedure leads to models that violate these conditions, we restart it with a different random seed until the condition is met. This happened in very rare cases (for details see Appendix C). Using these pairs of multilevel VAR models, we sample data from $N_{\text{subj}} \in \{50, 100, 200\}$ subjects, where either 20% or 50% of the subjects are in the first group.

### Estimating group differences

We estimate group differences using the R-package *mnet* (Haslbeck, 2023), which performs both the parametric and the permutation test described above using the R-package *mlVAR* (Epskamp et al., 2018). We estimate mlVAR models including random effects since part of the goal of the simulation study is to evaluate how well differences in random effects variances can be recovered. We specify that all random effects except random effects between intercepts (i.e., the between-network) are uncorrelated since the former are

rarely of interest and the run time would be much longer for correlated random effects.

Since the method implemented in the *mlVAR* package estimates multilevel VAR models in a stepwise manner that requires some estimates to be combined across nodewise regressions, we need to clarify on which exact parameters *mnet* performs inference. The lagged effects do not need to be combined across regressions and we perform inference directly on them. We perform inference on regression coefficients that are used to obtain estimates for the partial correlations between random intercepts, and the regression coefficients that are used to obtain estimates for the partial correlations between residuals after averaging them across nodewise regressions, but before normalizing them to partial correlations. We chose this approach because that way we could use the associated standard errors for the parametric test. This also involves averaging respective standard errors across nodewise regression, which we deem acceptable because for our standardized data the partial variances of all variables should be roughly the same which should also lead to similar standard errors.

We do not expect to reach perfect performance for recovering group differences between partial correlations between random intercepts (the "between-person" network). The reason is that the approximative method implemented in the *mlVAR* package gives estimates of the population between-person network that are biased to an extent that depends on the respective within-person correlations between variables (e.g., Haslbeck & Epskamp, 2023a; Hamaker, 2022).

The permutation test requires specifying a number of permutation. Any choice is necessarily arbitrary, but based on the reasoning in Appendix A we chose $N_P = 1000$ permutations.

### Evaluating performance

We evaluate simulation results using Receiver Operating Characteristic (ROC) curves (e.g., Fawcett, 2004), which plot the false positive rate (FPR) against the true positive rate (TPR). ROC curves are an attractive way to evaluate the performance of classifiers, because they allow us to investigate the performance of a method for the full range of values of a critical tuning parameter. This avoids that results depend on the arbitrary choice of a tuning parameter and also provides results for different choices of tuning parameters that offer different trade-offs between FPR and TPR. In our context, the tuning parameter is the significance threshold $\alpha$ which we vary between 0 and 1. For $\alpha = 1$ no group difference can be significant, which means that TPR = 0 and FPR = 1; on the other hand, for $\alpha = 0$ all group differences are significant and we get TPR = 1 and FPR = 0. Thresholds $\alpha$ chosen in practice will lie between those extremes and depend on the desired trade-off between false positives and false negatives.

To obtain estimates of the population performance, we run the design with $3\,(\Delta) \times 3\,(N_{subj}) \times 2\,(p_{G1}) \times 2\,(p) = 36$ cells for 50 repetitions. We chose 50 repetitions to strike a balance between the considerable cost in fitting many multilevel VAR models in the permutation test and obtaining precise performance estimates (see Appendix B). The code to reproduce the simulation study and all results in the paper can be found at https://github.com/jmbh/mlVARGD.

## Results

### Results for parametric test

Figure 1 shows the results of the parametric test as a function of the type of parameter (rows) and different sizes of group differences (columns). Note that there are only three types of parameters, fixed lagged effects, fixed effects between residuals, and partial correlations between random intercepts (the "between-network"), since *mlVAR* provides standard errors only for those parameters. In each panel, we display ROC curves, which are plotted on the plane of the false positive rate (FPR) on the $x$-axis and the true positive rate (TPR) on the $y$-axis. The diagonal line indicates the performance of an estimator that estimates group differences at random and we therefore expect the ROC curve of any reasonable method to lie above this line. We plot the results for different numbers of subjects in colors and use solid vs. dashed lines to differentiate between balanced (50:50) and unbalanced (20:80) groups.

We first consider group differences in lagged effects in the first row. We see that if group differences are very small ($\Delta = 0.04$) the performance in recovering the group differences is relatively low for all $N_{subj}$ and both balanced and unbalanced groups. For example, when fixing the alpha threshold at $\alpha = 0.05$, we get true positive rates (i.e., sensitivity) of 0.10, 0.24, and 0.46 for $N_{subj} = 50, 100, 200$ with balanced groups. For group differences with $\Delta = 0.08$ the performance is already much higher, leading to true positive rates of 0.46, 0.77, and 0.96 for $N_{subj} = 50, 100, 200$, when setting $\alpha = 0.05$. For the largest considered group differences with $\Delta = 0.12$ we see again a marked increase in performance. Here, we observe true positive rates of 0.76, 0.96, and 1.00 for $N_{subj} = 50, 100, 200$. Across variations of $\Delta$ and $N_{subj}$, the performance is lower for unbalanced groups. These overall results are very similar to the ones for testing group differences in fixed effects of innovation partial correlations.

If the test works perfectly, the specified $\alpha$ level aligns with the observed FPRs. For the three scenarios discussed above where we specified $\alpha = 0.05$, the observed average FPRs are

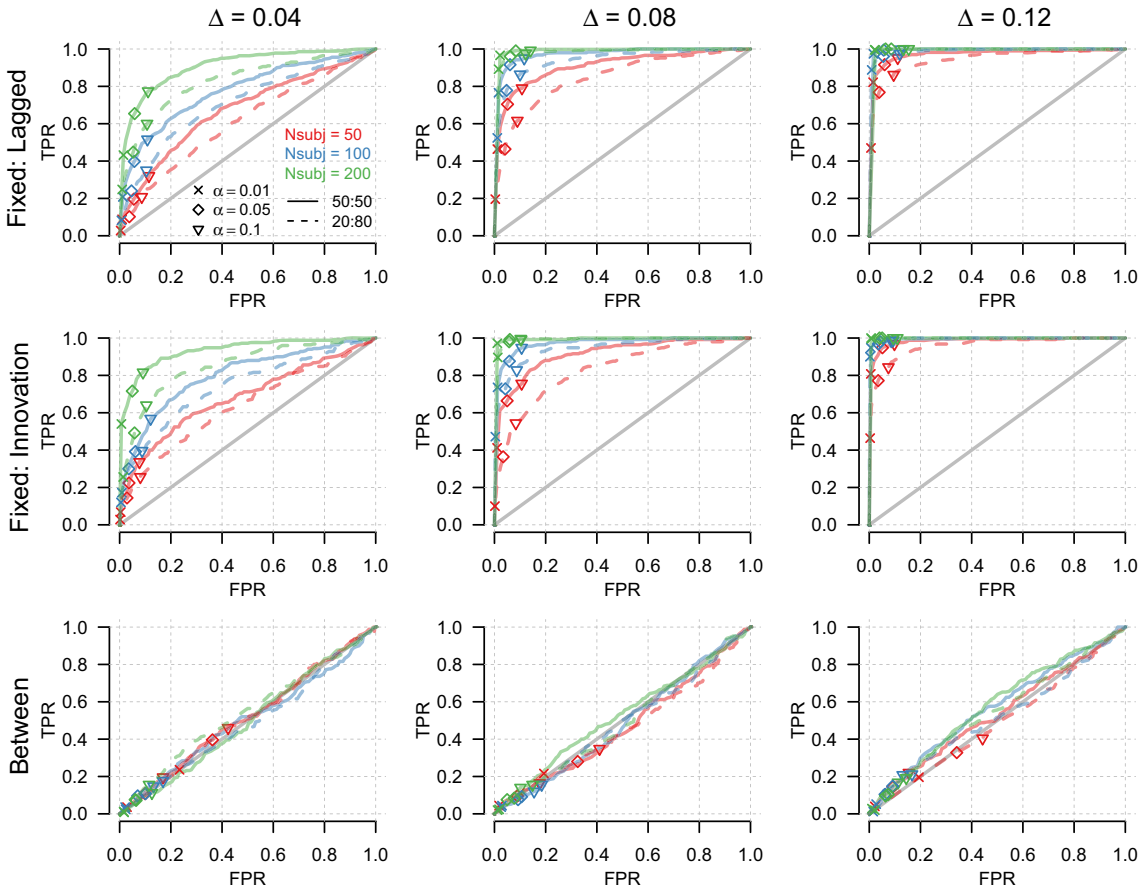


**Fig. 1** ROC curves for the parametric test for different numbers of total subjects (*colors*) and proportions in the first group (*dashed* vs. *solid lines*) for group differences of different sizes (*columns*) for the three variable types (*rows*) for which the parametric test could be performed. The *diagonal lines* indicate random performance. ROC curves plot the FPR (x-axis) against the TPR (y-axis) for a sequence of $\alpha$ values from 0 to 1. We in addition display the popular choces for $\alpha$ 0.01, 0.05 and 0.10. For standard errors of mean FPR/TPR values see Appendix B

0.04, 0.06 and 0.07 for $\Delta = 0.04$, the average FPR is 0.04, 0.05 and 0.06 for $\Delta = 0.08$, and the average FPR is 0.04, 0.06 and 0.07 for $\Delta = 0.12$. We see that the TPRs are somewhat inflated for larger $\Delta$ and $N_{subj}$, which we investigate in detail in Appendix D.

In contrast to the first two parameter types, the performance for recovering group differences in the "between-network" (bottom row) is extremely poor. Indeed, the ROC curves lie on the diagonal, which means that for this parameter the test is no better than determining group differences at random. As mentioned above, the reason is that the two-step procedure implemented in the *mlVAR* package leads to biased estimates of the population between-person network. This bias corrupts the detection of group differences in between-person networks in two ways: first, group differences in lagged effects and innovation correlations lead to different biases in the between-networks across groups, which will be picked up incorrectly as population group differences. Second, if within-person correlations between variables are high, they imply also large between-person correlations, which can lead to a ceiling effect that corrupts the detection of population differences in between-person networks. For more details on this bias, when it occurs, and possible solutions, see Haslbeck and Epskamp (2023a).

### Results for nonparametric permutation test

The results for the nonparametric permutation test are shown in Fig. 2. We again first focus on the fixed lagged effects in the first row. The overall results are very similar to the parametric test. For small group differences with $\Delta = 0.04$, fixing $\alpha = 0.05$ leads to TPRs of 0.12, 0.24, and 0.44 (FPRs: 0.05, 0.05, and 0.05) for $N_{subj} = 50$, 100, 200 with balanced groups. For $\Delta = 0.08$, the TPRs increase to 0.46, 0.76, and 0.96 (FPRs: 0.05, 0.05, and 0.06), and for $\Delta = 0.12$, the TPRs are 0.77, 0.95, and 1.00 (FPRs: 0.05, 0.05 and 0.05), respectively.

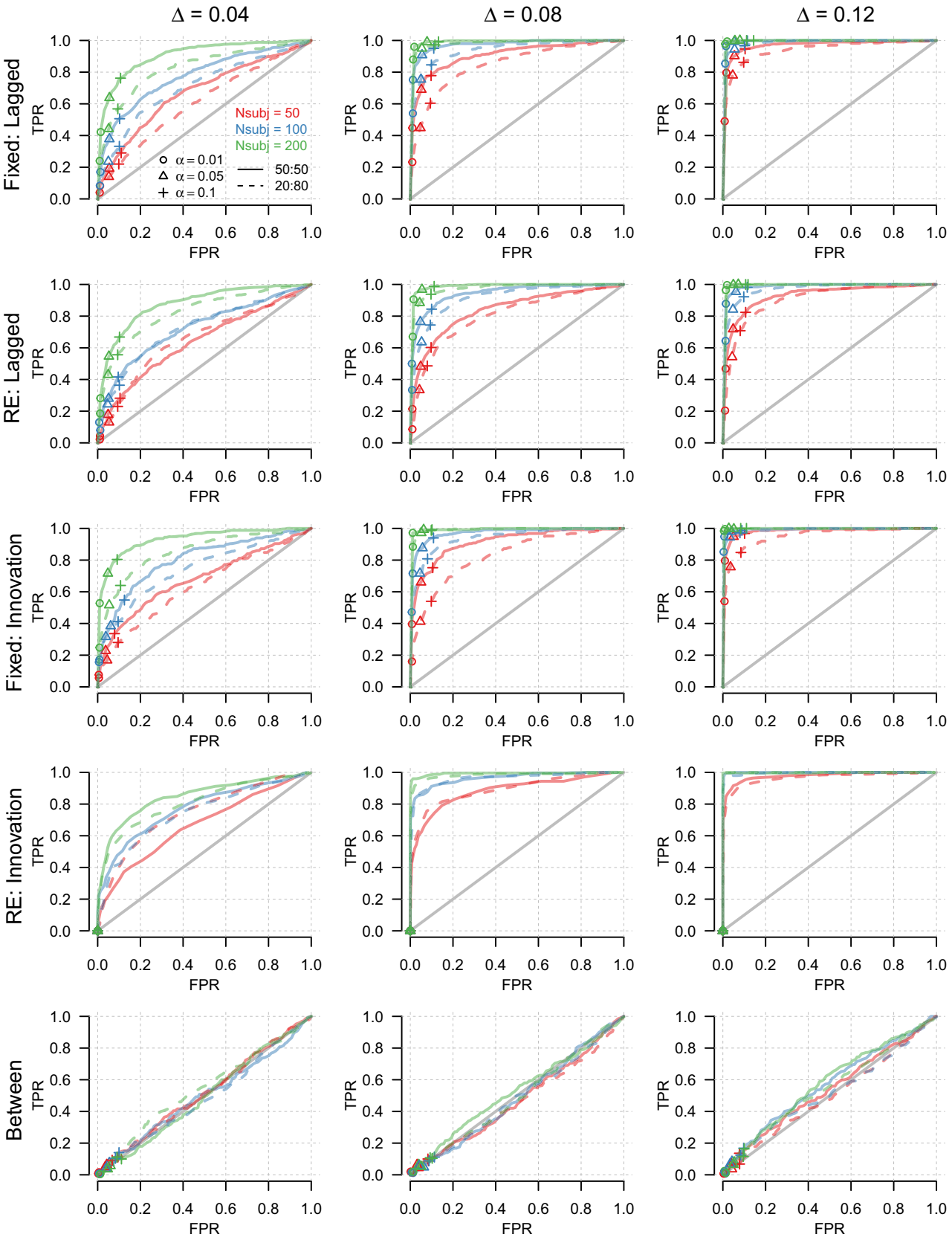


**Fig. 2** ROC curves of the nonparametric tests for different numbers of total subjects (*colors*) and proportions in the first group (*dashed* vs. *solid lines*) for group differences of different sizes (*columns*) for the five considered variable types (*rows*). The diagonal indicates random performance

The permutation test allows testing for group differences also in random effects variances of the lagged and the contemporaneous effects. We see that performance depends on the size of the group difference $\Delta$, $N_{subj}$ and group balance in the same way as the fixed lagged effects.

For the test for group differences in random effects variances in innovation partial correlations, we see that the three displayed $\alpha$ values are all at FPR = 0 and TPR = 0. This is an artefact stemming from the fact that the underlying algorithm often provides missing values for random effects variances close to zero. This in turn, downward-biases the variance of the sampling distribution under the null hypothesis of the permutation test, which corrupts the test. We discuss this issue in detail in Appendix E and return to it in the discussion.

### Comparing results of both tests

To better compare the parametric and the permutation test, we compare ROC curves from both tests in Fig. 3. We do not compare the performance for the between-network since the performance is at random levels for both tests. For group differences in fixed lagged effects and fixed innovation partial correlations, we see that the ROC curves are essentially identical.

However, fixing a specific threshold $\alpha$ we see that the performance differs in the case of group differences in fixed lagged effects. Specifically, we see that that both the TPR and the FPR of the parametric test are higher than for the nonparametric permutation test, indicating that the parametric test is more liberal and that the FPRs are more inflated for this test. This is also in line with the TPRs and FPRs reported for $\alpha \in \{0.01, 0.05, 0.10\}$ above. We investigate and discuss the alignment between observed FPRs and specified $\alpha$ thresholds in more detail in Appendix D.

### Discussion of results

In the previous section, we have seen that both the parametric test and the nonparametric test perform very similarly in detecting group differences for the type of parameters that allow a comparison. In addition, we saw that the performance is relatively low for $\Delta = 0.04$, meaning that if one chooses a typical $\alpha$ threshold such as 0.05, the TPR (sensitivity) will be relatively low. For larger group differences of $\Delta = 0.08$, high performance can be achieved with $N_{subj} = 200$, and for group differences with $\Delta = 0.12$ high TPRs are possible even with small sample sizes, while fixing FPRs at typical levels. We saw that the performance was similar across different parameter types. The exceptions are that detecting group differences in the fixed innovation partial correlations was somewhat easier than detecting group differences in fixed lagged effects. Furthermore, the fact that recovering group differences in the correlations between random intercepts (i.e., the between-persons network) was not possible at all due to biases in the two-step procedure used in *mlVAR* (Haslbeck & Epskamp, 2023a).

While the ROC curves were similar across the two tests, we saw that the behavior was slightly different for fixed significance thresholds $\alpha$. We saw that the FPR aligned well with the theoretical $\alpha$ for small group differences $\Delta$, but there was some misalignment for large $\Delta$ and large $N_{subj}$. We further explore and discuss this misalignment in Appendix D. We

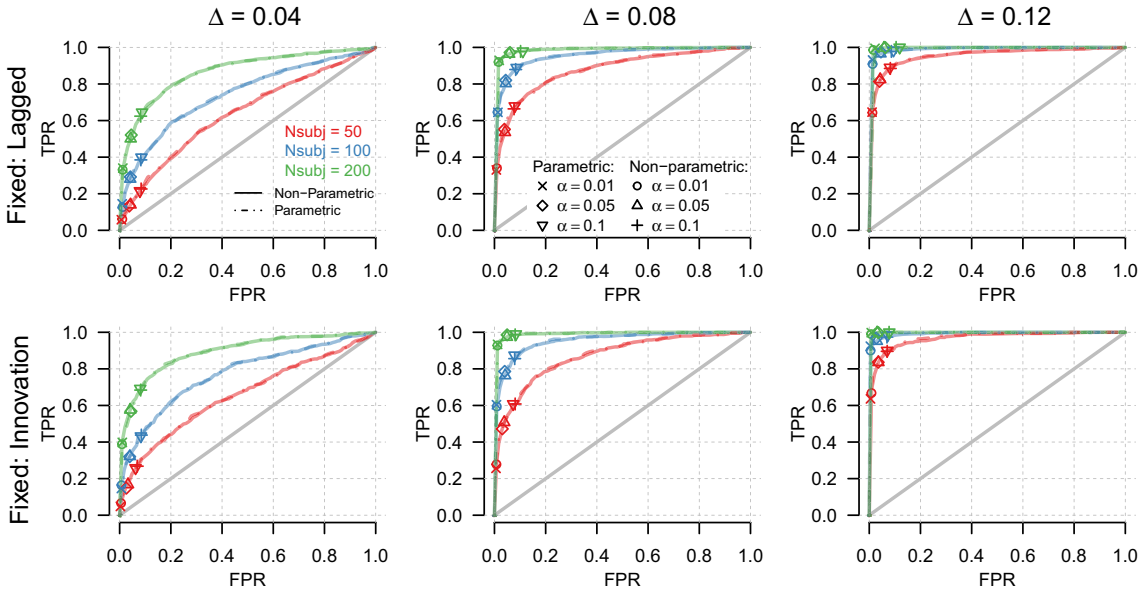


**Fig. 3** ROC curves for different numbers of total subjects (*colors*) and type of test (*solid lines* = permutation test; *dashed lines* = parametric test) for group differences of different sizes (*columns*) for the five considered variable types (*rows*). The diagonal indicates random performance

expect that this misalignment is due to the two-step procedure employed by the *mlVAR* package and that performing the tests using estimates from a joint multilevel model such as in DSEM (Asparouhov et al., 2018), the $\alpha$ values would line up with expected FPR also for fixed lagged effects.

Based on these results, which method should be preferred in empirical research? We simulated data in a scenario where both the innovation variances and the heterogeneity across individuals were sampled from a Gaussian distribution, which contributes to a relatively fast convergence of sampling distributions to approximately Gaussian distributions, as is assumed by the parametric test. Despite this fact, we have seen that the permutation test does not perform worse than the parametric test. In empirical research, we are likely to encounter situations where innovations and heterogeneity across individuals are not Gaussian distributed, which would negatively impact the parametric test more than the permutation test. This suggests that the permutation test should be preferred in practice, at least in scenarios similar to those studied in our simulation. Additional reasons for preferring the permutation test include its ability to test group differences in random effects variances. Furthermore, the permutation test facilitates group comparisons in within-subject designs, such as evaluating a group before and after an intervention. The downside of the permutation test is that it is computationally intensive.

### Limitations

As in any simulation study, we had to keep various parameters fixed, which limits the range of our conclusions. For example, we kept the number of variables $p$, the number of time points per subject $N_t$, and the random effects variances constant. While we chose fixed parameters to best represent typical empirical settings, strictly speaking our results only hold for the conditions we simulated. However, we can understand all (fixed) parameters as contributing to the signal-to-noise ratio (SNR) associated with group differences. This means that, for example, increasing $N_{\text{subj}}$ and decreasing $\Delta$ leads to the same SNR and therefore the same results. In addition, we did not correlate the random effects of lagged and contemporaneous effects. However, as long as the model is correctly specified, we do not expect that estimation errors are larger for correlated random effects. Consequently, if random effects are correlated in the true model and random effects correlations are estimated, the estimation error remains the same and therefore also the performance in recovering group differences is similar.

## R-tutorial: Application to emotion time series data

We use data from the Ecological Momentary Assessment (EMA) study of Koval, Pe, Meers, and Kuppens (2013) which consists of measures administered up to 10 times a day for 7 days of 95 undergraduate students at the KU Leuven in Belgium. The measures included ratings of the seven emotion variables *Happy*, *Relaxed*, *Sad*, *Angry*, *Anxious*, *Depressed*, and *Stressed* since the last measurement occasion. The seven emotion variables were scored on a 0-100 Visual Analog Scale. All subjects were assessed with the CES-D (Radloff, 1977) to measure current depressive symptoms, which includes 20 items that are scored on a 0–3 scale. We use the clinical cutoff of the CES-D $\leq$ 16 suggested by Radloff (1977) to create a group with low (62 subjects) and high (33 subjects) depressive symptoms.

These data are automatically loaded with the *mnet* package, which is available from the Comprehensive R Archive Network (CRAN):

```
library(mnet)
dim(dataKoval13)
[1] 3908   14

head(dataKoval13)
  PID UNIT OCCASION   PA  NA. Happy Relaxed Sad Angry Anxious Depressed Stressed CESD group
1   2   25        1 66.0  4.0    49      83  12     1       1         5       1 0.45     1
2   2   25        2 35.5 12.8    31      40  20    21       1        21       1 0.45     1
3   2   25        3 36.0 30.4    20      52  58     1      22        38      33 0.45     1
5   2   25        5 77.0 13.0    73      81   1     1       1         1      61 0.45     1
6   2   26        6 62.0  1.0    41      83   1     1       1         1       1 0.45     1
7   2   26        7 60.5  4.2    40      81   1     1       1         1      17 0.45     1

> length(unique(unique(dataKoval13$PID)))
[1] 95
```

We see that the data frame contains 14 variables, which include a participant id (`PID`), the day number (`UNIT`), the measurement occasion within the day (`OCCASIOON`), the group indicator (`group`) which we have computed from the CES-D scale (see above). The data also include measures for positive and negative affect, which are aggregates of the seven emotion variables and which we will not use in the below analysis. We also see that there are a total of 3908 rows in the data frame and that there are data from 95 subjects.

To test differences with high and low CES-D scores, we provide the seven emotion variables to the function `mlVAR_GC()`, indicate with the argument `vars` which of the variables should be included in the model, and with the arguments `id`, `dayvar`, `beepvar` and `groups` the column names of variables indicating unique identifiers for subjects, information on the measurement day and occasion, and the group membership. We specify that we would like to perform the permutation test with `test = "permutation"` and set the number of perturbations to $N_P = 1000$, the number of cores to 12, and request with `pbar = TRUE` that a progress bar is being shown. By default, a VAR model with a lag of 1 is fit to the data.

```
output <- mlVAR_GC(data = dataKoval13,
                   vars = colnames(dataKoval13)[6:12],
                   idvar = "PID",
                   dayvar = "UNIT",
                   beepvar = "OCCASION",
                   groups = "group",
                   test = "permutation",
                   paired = FALSE,
                   contemporaneous = "orthogonal",
                   temporal = "orthogonal",
                   nP = 1000,
                   nCores = 12,
                   pbar = TRUE)
```

The runtime was 89 min on 12 cores of a 2.6 GHz Intel i7 and can be found in minutes in `output$Runtime_min`. The empirically observed group differences for the different parameter types can be found in `output$EmpDiffs` and the corresponding $p$ values can be found in `output$Pvals`. The two multilevel VAR models estimated on the empirical data can be found in `output$ModelsEmp`. To keep the tutorial short, we focus only on group differences in the fixed temporal effects. We plot the fixed lagged effects for the two groups, the observed group differences and the group differences that are significant with $\alpha = 0.05$ based on the permutation test in Fig. 4. The complete code to reproduce the figure from the output of `mlVAR_GC()` can be found at https://github.com/jmbh/mlVARGD.

The top panels of Fig. 4 show the fixed lagged effects for the low and high CES-D groups. A visual inspection of these two VAR networks and the network of differences in the bottom left panel suggests that the networks of individuals with high and low CES-D scores are substantially different. However, if we consider only the group differences that were significantly different at threshold $\alpha = 0.05$, we see that only three group differences remain. This highlights the importance of doing statistical inference to infer group differences, rather than using eye-balling or descriptive statistics alone. Since estimating group differences in parameters is based on estimating parameters, these results are in line with research showing that accurate parameter estimation of VAR models often requires more time points than are available in many ESM-studies (Mansueto et al., 2022; Dablander et al., 2020; Bulteel et al., 2018).

How certain are we about the presence and absence of group differences in this setting? Taking the estimated significant group differences as estimates of true group differences, we have a $\Delta \approx 0.10$, $N_{subj} = 95$ which which is close to the simulated condition with $N_{subj} = 100$, and we have an unbalanced group. Looking at the middle panel of the first row in Fig. 2 we see that, for threshold $\alpha = 0.05$, we expect a TPR of around 0.90 and a FPR of around 0.05.

## Discussion

We introduced a parametric and a nonparametric permutation test to make inferences about group differences in the parameters of multilevel Vector Autoregressive (VAR) models. We discussed the rationale for the tests and discussed their advantages and disadvantages based on theory. We then evaluated the performance of both tests in scenarios resembling those in which multilevel VAR models are typically applied.

We found that moderate group differences ($\Delta = 0.08$) can already be recovered relatively reliably (TPR $\approx 0.74$, FPR $\approx 0.05$) with 200 subjects, and for larger group differences ($\Delta = 0.12$), a similar performance can already be achieved with 100 subject. We found that unbalanced groups (20:80 vs. 50:50) considerably reduced performance, but this difference was smaller for larger group differences and more subjects, where the overall performance was higher. The results were similar for all parameter types except for the between-person network, for which performance was extremely low due to a known bias of the estimation method used in the *mlVAR* package. Finally, we provided a fully reproducible tutorial showing how to use the R-package *mnet* to differences in emotion dynamics across groups with low vs. high depressive symptoms.

The presented tests allow researchers to make inferences about group differences for the fixed effect (parametric test) and both the fixed and random effects variances (permutation test) of lagged effects and innovation partial correlations. However, the performance of both tests to recover group differences in the between-network was extremely low. This is due to biased between-person network estimates of the

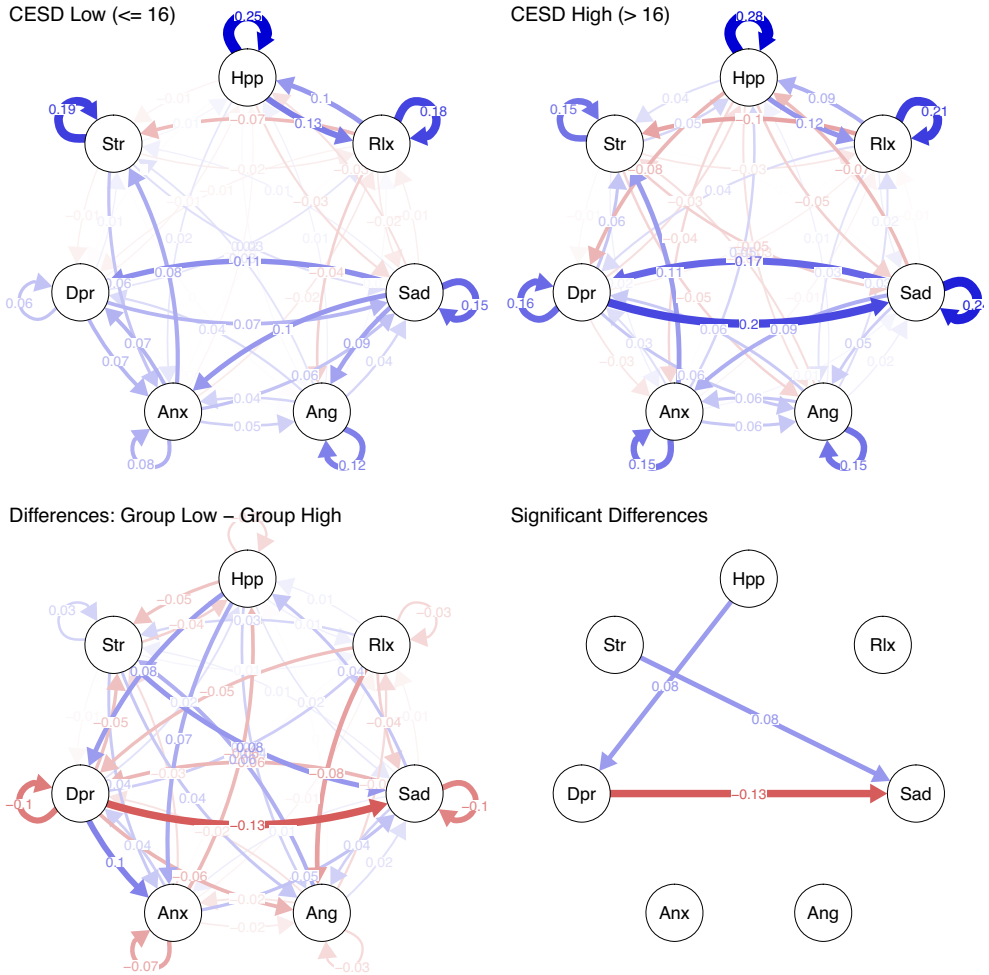


**Fig. 4** Inspecting group differences in fixed lagged effects in the data of Koval et al. (2013). Hpp = Happy, Rlx = Relaxed, Sad = Sad, Ang = Angry, Anx = Anxious, Dpr = Depressed, Str = Stressed; The *top panel* shows the fixed lagged effects of the two groups; the *bottom left panel* shows group differences (Group Low - Group High); and the *bottom right panel* shows the group differences that are significant with $\alpha = 0.05$ based on the permutation test (*bottom right*)

method implemented in the *mlVAR* package (for details, see Haslbeck & Epskamp, 2023a). To test group differences in between-person networks this bias has to be avoided, for example by estimating the mlVAR model in a single step such as in the DSEM module of M*plus* (Asparouhov et al., 2018; McNeish & Hamaker, 2020). However, DSEM currently does only provide correlations and not partial correlations. While correlations can be transformed into partial correlations, we are not able to obtain standard errors on the latter using DSEM.

When evaluating the parametric and the permutation test for recovering group differences in fixed lagged effects, we found that the FPRs were slightly inflated. This inflation becomes larger for larger group differences. We traced this outcome back to larger variances of sampling distributions of group differences that are absent in the true model, when the present group differences become larger. We were not able to find a single explanation of this effect, but instead found that the effect was dependent on many specifications of the multilevel model. Our hypothesis is that this effect is due to misspecification arising from the two-step approach employed by the *mlVAR* method, which we used to estimate multilevel VAR models. For applied researchers, this means that the FPRs are inflated as mapped out in our simulation study, which should be taken into account when reporting results.

Based on the results of the simulation study we recommend using the permutation test in empirical research. The reason is that the setting of the simulation study was designed to favor the parametric test since we assumed Gaussian innovations and between-person heterogeneity in empirical data. Despite this theoretical advantage for the parametric test, we found that the performance of the two tests was very similar for the types of parameters that could be compared.

This suggests that the permutation test should be preferred because it is less affected by violations of Gaussian innovations and heterogeneity. In addition, the permutation test allows one to test group differences in random effects variances and allows group comparison across paired samples. Finally, the permutation test allows researchers to test for group differences within individuals at different time points. The downside of the permutation test is its much longer runtime, since it involves estimating a large number of multilevel VAR models.

An alternative way to test group differences would be to dummy-code group membership and introduce interaction terms between all parameters and these dummy-coded level 2 predictors. This approach is similar to the one Haslbeck (2020) used to test group differences in cross-sectional network models. An advantage of this approach is that it allows to do inference about differences across more than two groups in a single model, which can subsequently be compared using tests between estimated marginal means. This approach was used by Bringmann et al. (2013) to test for differences between groups in multilevel VAR models. However, this approach has not yet been extended to separate between within- and between-person effects as well as to estimate contemporaneous effects, and is not implemented in the *mlVAR* package.

The group comparison studied here can be seen as testing a moderation effect with a binary variable. This approach is most appropriate if the grouping variable is nominal or ordinal with two categories. However, in many situations moderators are ordinal with more than two categories or continuous. For example, in the tutorial we created the grouping variable by binarizing the scores on the CES-D scale, which ranges between 0 and 60. Any parameter in the multilevel VAR model can be moderated by the CES-D score and this moderation can in principle take any functional form. When binarizing the score at the mean, we are implicitly assuming that this functional form is well approximated by a step function with the step at cutoff we used for binarization. However, often there is no good reason to make such assumptions a priori. In such settings, methods that estimate multilevel VAR parameters as continuous functions of the moderator variable, for example using kernel-smoothing (e.g., Hastie et al., 2009) or generalized additive models (e.g., Wood, 2017), would be preferable. Methods to estimate such moderation effects have been developed in the context of time-varying fixed effects VAR models (Haslbeck, Bringmann, & Waldorp, 2021), which could also be used for moderators other than time. Extending these methods to a multilevel context would be a useful avenue for future research.

In most situations, the permutation test requires a considerable amount of computational cost, which can limit its usefulness in some settings. One way to make permutation tests computationally more efficient is to approximate its tails with a parametric distribution. This reduces computational cost significantly, because $p$ values in the tails are the ones that require the most permutations to be estimated accurately (see Appendix A). For example, Knijnenburg, Wessels, Reinders, and Shmulevich (2009) show that the required permutations can be reduced considerably by approximating the tails of sampling distributions with generalized Pareto distributions fitted to extreme values. An interesting direction for future work would be to examine to what extent such approaches can reduce the number of permutations in the present context.

In summary, we have introduced two tests for group differences in multilevel VAR models, provided implementations in the new R-package *mnet*, used a simulation study to evaluate the methodology in situations similar to empirical research, and provided a fully reproducible tutorial on how to use the method to test for differences in emotion dynamics across groups with high vs. low depressive symptoms. We hope that this methodology will help empirical researchers to make more reliable inferences about group differences in multilevel VAR models.

## A. Choosing the number of permutations

The computational cost of estimating a multilevel VAR model is nontrivial and therefore it is important to carefully consider how many permutations are required to obtain good estimates of sampling distributions and $p$-values. For example, estimating a pair of multilevel VAR models with orthogonal random effects, $p = 6$ variables and $N_{\text{subj}} = 100$ subjects per group with the mlVAR package takes around 1.2 min on a 2.6 GHz Intel i7 core. Therefore, running $N_{\text{P}} = 1000$ permutations sequentially would already take more than 21 h. We therefore need to carefully consider how many permutations are required for accurate inferences about group differences. Here, we will discuss how many permutations are necessary, drawing both on statistical theory and simulations in the context of multilevel VAR models.

We will use the sampling distributions under the null hypothesis to compute $p$ values for observed test statistics. We therefore want those $p$ values to be accurately estimated by choosing a sufficient number of permutations. Theoretical work on permutation tests (e.g., Tibshirani & Efron, 1993, p.209) specifies the standard deviation of a specific *estimated* $p$ value $\hat{P}$, under the null hypothesis, obtained with

the permutation test as a function of the *population* $p$ value $P$ and the number of permutations $k$:

$$\mathrm{sd}(\hat{P}) = \sqrt{\frac{P(1-P)}{k}} \tag{1}$$

This is the same as the standard error of a Bernoulli random variable. This shows that the accuracy is highest for $P$ close to 0 and 1 and lowest for $P = 0.5$. In addition, we see that the standard deviation decreases with $\sqrt{k}$. For example, if we would consider a standard deviation $\mathrm{sd}(\hat{P}) = 0.005$ acceptable and the true $p$ value $P$ is equal to 0.05, then we need $k \geq \frac{P(1-P)}{\mathrm{sd}(\hat{P})^2} = \frac{0.05(1-0.05)}{0.005^2} = 1900$ permutations. On the other hand, if the $p$ value is extremely small, only very few permutations are needed: $k \geq \frac{P(1-P)}{\mathrm{sd}(\hat{P})^2} = \frac{0.001(1-0.001)}{0.005^2} \approx 40$. This shows that if we have very low uncertainty about a group difference, then already very few permutations are sufficient. However, to estimate $p$ values in the segment of the tail of the sampling distribution in which critical cutoff values are typically located (e.g., 0.05 leading to $k \geq 1900$ or 0.01 leading to $k \geq 396$), more permutations are required. We provide more intuition about the permutation test and the

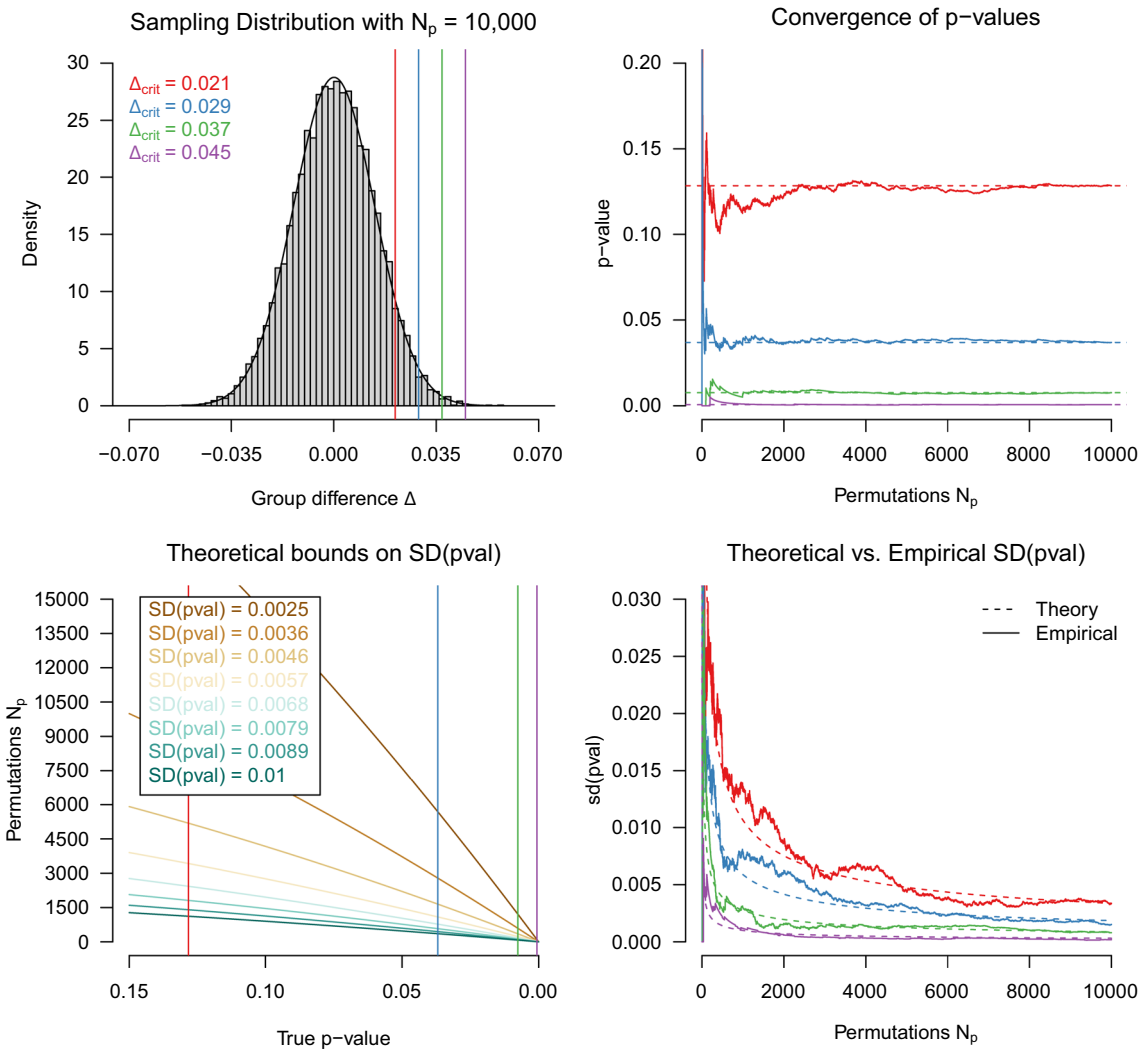


**Fig. 5** *Top left*: histogram of sampling distribution obtained with $N_p$ = 10,000 permutations. We choose four critical values that are associated with $p$ values close to typical thresholds like 0.05 and 0.01; *Top right*: estimated $\hat{P}$ associated with the four critical values as a function of $N_p$; *Bottom left*: the required number of permutations $k$ ($y$-axis) as a function of the true $p$ value $P$ ($x$-axis) and the desired accuracy $\mathrm{sd}(\hat{P})$; *Bottom* right: comparison of the theoretical and empirical $\mathrm{sd}(\hat{P})$ as a function of the true $p$ value (*colors*) and $N_p$ ($x$-axis)

required number of permutations to obtain good estimates for population sampling distributions and $p$ values in Fig. 5.

The top left panel of Fig. 5 shows a histogram of the sampling distribution we obtained with $N_P = 10,000$ permutations. Since we used simulated data in which group differences are generated from a Gaussian distribution, we see that the histogram of the sampling distribution closely fits a Gaussian density. However, the permutation test also works for most other underlying distributions, as specified by the Central Limit Theorem. We consider four critical values 0.021, 0.029, 0.037, and 0.045 in the tail of the sampling distribution where also the typical significant thresholds of 0.05 or 0.01 are located. Smaller critical values are not too interesting because even relatively large errors on associated $p$ values would not change the outcome of the test. And much larger critical values would lead to extremely precise estimates, because we will estimate $\hat{P} = 0$ since we have no support further out in the tails in the empirical sampling distribution.

In the top right panel we display the $p$ values associated with these four critical values as a function of the number of permutations $N_P$. In line with the theoretical result in Eq. 1, the small $p$ value associated with $\Delta_{crit} = 0.045$ converges extremely quickly. In contrast, the largest considered $p$ value associated with $\Delta_{crit} = 0.021$ takes much longer, only stabilizing around $N_p = 4{,}000$. The bottom left panel displays the theoretical $N_P$ required to achieve different $\text{sd}(\hat{P})$ for the range of most relevant $p$ value between 0.1 and 0. This again illustrates the fact that if a group difference is large, and the corresponding $p$ value is correspondingly small, already a very small amount of permutations are sufficient for precise $\hat{P}$ and accurate inferences.

Finally, in the bottom right panel we compare both the theoretical and estimated $\text{sd}(\hat{P})$ for the four critical values as a function of $N_p$. For the empirical demonstration, we simulated data from the multilevel VAR model we also use in the simulation reported in the main text. However, as is also clear from Eq. 1, the results are independent from the specifics of the data-generating process. For example, if we had fewer subjects per group or larger innovation variances, then the variance of the sampling distribution in the top left panel would be much wider. While this means that different critical values will now be associated with given $p$ values, this does not interact with how many $N_P$ are required to estimate a given $p$ value with a desired precision. For the same reasons, the results above extend to the sampling distributions under the null hypothesis for other parameters such as random effects variances.

So, how many permutations are required for the permutation test for group differences in multilevel VAR models? As is clear from the theoretical result and the demonstration with simulated data, the accuracy of $p$ values are a monotone function of $N_P$ and choosing a cutoff is necessarily arbitrary. However, when focusing on $p$ values around significance thresholds of $\alpha = 0.05$ then $N_p > 1000$ already allow for relatively accurate inferences about group differences.

## B. Uncertainty about mean estimates of reported FPR/TPR

We limited the repetition of the simulation design to 50 since the computational cost of the permutation test for the multilevel VAR model was considerable. Specifically, a single repetition of the design takes about 288h to run on a single core of the 2.6 GHz CPUs of the Dutch national supercomputer Snellius. We expected stable results nonetheless, since we report FPRs and TPRs which are already aggregates across $8^2 = 64$ edges for the fixed lagged parameter and associated random effects variances and $\frac{8\times7}{2} = 28$ for all other types of parameters.

To get a measure for the precision of the estimates in our simulation study, we compute the standard error of the mean for all results shown in Figs. 1 and 2. The size of the group difference and whether the groups were balanced did not have a meaningful influence on the standard error and we therefore averaged over those two dimensions in the tables below. Tables 1 and 2 display the resulting aggregate standard errors

**Table 1** Standard deviations and standard errors for TPRs and FPRs of the parametric test in the main text

| #Subjects | Parameter Type | Outcome | SD | SE |
|---|---|---|---|---|
| 50 | Fixed: Lagged | TPR | 0.328 | 0.046 |
| 100 | Fixed: Lagged | TPR | 0.313 | 0.044 |
| 200 | Fixed: Lagged | TPR | 0.233 | 0.033 |
| 50 | Fixed: Innovation | TPR | 0.346 | 0.049 |
| 100 | Fixed: Innovation | TPR | 0.327 | 0.046 |
| 200 | Fixed: Innovation | TPR | 0.239 | 0.034 |
| 50 | Between | TPR | 0.248 | 0.035 |
| 100 | Between | TPR | 0.138 | 0.020 |
| 200 | Between | TPR | 0.128 | 0.018 |
| 50 | Fixed: Lagged | FPR | 0.032 | 0.004 |
| 100 | Fixed: Lagged | FPR | 0.031 | 0.004 |
| 200 | Fixed: Lagged | FPR | 0.038 | 0.005 |
| 50 | Fixed: Innovation | FPR | 0.041 | 0.005 |
| 100 | Fixed: Innovation | FPR | 0.044 | 0.006 |
| 200 | Fixed: Innovation | FPR | 0.047 | 0.007 |
| 50 | Between | FPR | 0.209 | 0.030 |
| 100 | Between | FPR | 0.069 | 0.010 |
| 200 | Between | FPR | 0.063 | 0.009 |

SD = standard deviation; SE = standard error

**Table 2** Standard deviations and standard errors for TPRs and FPRs of the permutation test in the main text

| #Subjects | Parameter Type | Outcome | SD | SE |
|---|---|---|---|---|
| 50 | Fixed: Lagged | TPR | 0.318 | 0.045 |
| 100 | Fixed: Lagged | TPR | 0.311 | 0.044 |
| 200 | Fixed: Lagged | TPR | 0.236 | 0.033 |
| 50 | RE: Lagged | TPR | 0.253 | 0.036 |
| 100 | RE: Lagged | TPR | 0.300 | 0.042 |
| 200 | RE: Lagged | TPR | 0.251 | 0.036 |
| 50 | Fixed: Innovation | TPR | 0.341 | 0.048 |
| 100 | Fixed: Innovation | TPR | 0.323 | 0.046 |
| 200 | Fixed: Innovation | TPR | 0.230 | 0.032 |
| 50 | RE: Innovation | TPR | 0.000 | 0.000 |
| 100 | RE: Innovation | TPR | 0.000 | 0.000 |
| 200 | RE: Innovation | TPR | 0.000 | 0.000 |
| 50 | Between | TPR | 0.119 | 0.017 |
| 100 | Between | TPR | 0.112 | 0.016 |
| 200 | Between | TPR | 0.109 | 0.015 |
| 50 | Fixed: Lagged | FPR | 0.031 | 0.004 |
| 100 | Fixed: Lagged | FPR | 0.030 | 0.004 |
| 200 | Fixed: Lagged | FPR | 0.036 | 0.005 |
| 50 | RE: Lagged | FPR | 0.029 | 0.004 |
| 100 | RE: Lagged | FPR | 0.032 | 0.004 |
| 200 | RE: Lagged | FPR | 0.031 | 0.004 |
| 50 | Fixed: Innovation | FPR | 0.040 | 0.006 |
| 100 | Fixed: Innovation | FPR | 0.044 | 0.006 |
| 200 | Fixed: Innovation | FPR | 0.048 | 0.007 |
| 50 | RE: Innovation | FPR | 0.000 | 0.000 |
| 100 | RE: Innovation | FPR | 0.000 | 0.000 |
| 200 | RE: Innovation | FPR | 0.000 | 0.000 |
| 50 | Between | FPR | 0.071 | 0.010 |
| 100 | Between | FPR | 0.050 | 0.007 |
| 200 | Between | FPR | 0.055 | 0.008 |

SD = standard deviation; SE = standard error

of FPR and TPR for fixed $\alpha = 0.05$ for the parametric and permutation test, respectively. For example, we see that, for fixed lagged effects, the standard error for the TPR is 0.044 and standard error for the FPR is 0.004 for $N_{subj} = 100$. Doubling the number of iterations in the simulation design would decrease standard errors by a factor of only $\frac{\sqrt{50}}{\sqrt{100}} \approx 0.707$, which to us does not justify doubling the computational cost.

## C. Resampling in data generation step of simulation

We generate data from the true models as specified in Section "Defining true models & data generation" in the main text. However, this specification does not guarantee that the VAR parameter matrix is stable (i.e., has eigenvalues in the unit circle), has no oscillation (only real eigenvalues) and that the innovation covariance matrix is positive-definite. Note that these requirements need to be enforced both for fixed effects and for random effects.

**Table 3** The average resamples and maximum resamples across iterations, for the two matrices (VAR matrix and innovation covariance matrix), separately for all simulation conditions

| Parameter Type | Delta | Mean (Resamp) | (Max Resamp) |
|---|---|---|---|
| VAR | 0.04 | 18 | 103 |
| VAR | 0.08 | 25 | 138 |
| VAR | 0.12 | 30 | 177 |
| Innovation | 0.04 | 0 | 4 |
| Innovation | 0.08 | 1 | 4 |
| Innovation | 0.12 | 1 | 6 |

We resample both when defining the fixed effects model and for each individual model drawn from the random effects distribution. Table 3 shows the mean and maximum number of resamples when specifying the fixed effects model and the random effects models. The numbers are rounded to the nearest integer. We see that most of the resamples were required to specify a valid fixed effects model. Specifically, these resamples are required to satisfy the conditional that the lagged effects matrix has eigenvalue within the unit circle.

The details of this implementation can be found in our reproducibility archive https://github.com/jmbh/mlVARGD. A possible concern with this procedure could be that if many resamples are necessary to satisfy those two conditions, we might bias our data-generating models in a direction that is not apparent from the description in Section "Defining true models & data generation". However, this is not the case since we require relatively few resamples overall.

## D. Alignment of specified $\alpha$ with observed FPR

In Figs. 1 and 2 in the main text, we indicated TPRs and FPRs of the popular choices of $\alpha \in \{0.01, 0.05, 0.1\}$ and which showed that FPRs were somewhat inflated, especially for the parametric test. Here, we investigate this inflation more systematically.

In Fig. 6 we plot the specified $\alpha$ level ($x$-axis) against the observed FPRs for the parametric test ($y$-axis), for the different conditions in our simulation study. If the test functions perfectly, the two are identical and fall on the diagonal. For fixed lagged effects, we see that while the FPRs are very close to the specified $\alpha$s, there seems to be an inflation that grows with larger group differences $\Delta$ and number of subjects $N_{subj}$. In fact, there seems to be an interaction between $\Delta$ and

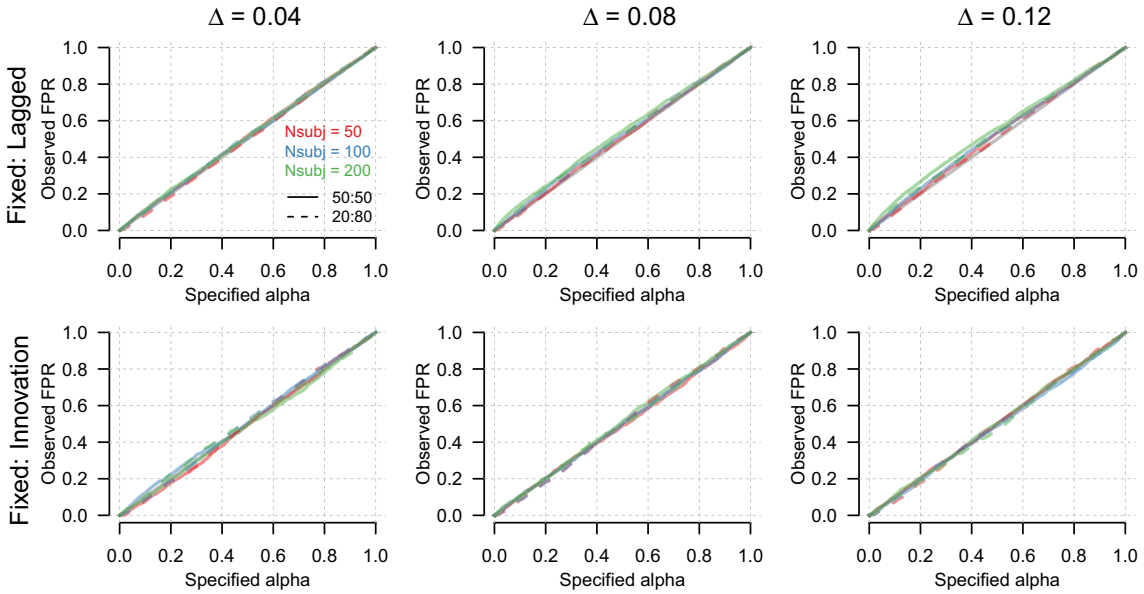


**Fig. 6** Specified $\alpha$-values ($x$-axis) plotted against the observed FPR in the simulation study, separately for all simulation conditions for the parametric test

$N_{subj}$. For fixed innovation partial correlations, the inflation is much smaller and difficult to distinguish from sampling variation given the number of iterations in our simulation. Before investigating this inflation more, we inspect the corresponding alignment for the permutation test.

Figure 7 shows the same analysis for the permutation test, albeit here we can also inspect the alignment for group differences in random effects variances. We see the same qualitative results for the fixed lagged effects, where the inflated FPRs seem to be a function of the size group differences $\Delta$ and number of subjects $N_{subj}$. We do not see a systematic inflation for the other parameters. We see an extreme deflation for the random effects variances of the innovation partial correlations. This behavior is an artifact we mentioned in the main text and we explain it in detail in Appendix E.

In what follows, we investigate the inflation of FPRs for the fixed lagged effects. Since both tests are similarly affected, we expect that the reason is not due to the specifics of either of the two tests. A false positive occurs if we determine a group difference for a parameter, even though there is no group difference in the population model. Therefore, we investigated the sampling distribution of the group differences for those parameters for which the group difference is zero in the population model to investigate how they are affected by the variations of the size of the group differences $\Delta$ and number of subjects $N_{subj}$.

Figure 8 shows the sampling distribution of the differences for those fixed lagged effects for which no group differences were present in the population model from which we generated the data. For any fixed group difference $\Delta$, we see that the spread of the sampling distributions becomes smaller, as one would expect when having more data to estimate the fixed effect. However, when keeping $N_{subj}$ fixed and increasing $\Delta$, we see that the spread of the sampling distribution increases. This increase is larger for higher $N_{subj}$, showing that the spread of the sampling distribution is a function of an interaction between $\Delta$ and $N_{subj}$.

If the increased spread of the sampling distribution for higher values of $\Delta$ and $N_{subj}$ is not accounted for in the tests, this would lead to higher FPRs especially for high $\Delta$ and $N_{subj}$, which is precisely what we see in Figs. 6 and 7. The parametric test constructs the sampling distribution under the null, in which group differences are by definition equal to zero, which means that the spread of the sampling distribution will be smaller than it should be, therefore explaining the inflated FPRs we observed. Similarly, the theory-derived spread of the sampling distribution under the null hypothesis is also lower than the actual sampling distribution under the null hypothesis we observe in our simulation study, leading to inflated FPRs.

We did not expect that increasing the group differences $\Delta$ would increase the spread of the sampling distribution of the differences between parameters for which population group differences were zero. We further investigated this phenomenon from two angles. First, we wanted to know whether this phenomenon is induced by group differences that are not in the fixed lagged effects. We therefore reran a part of our simulation that only included group differences in the fixed lagged effects. The inflation of FPRs was now much smaller

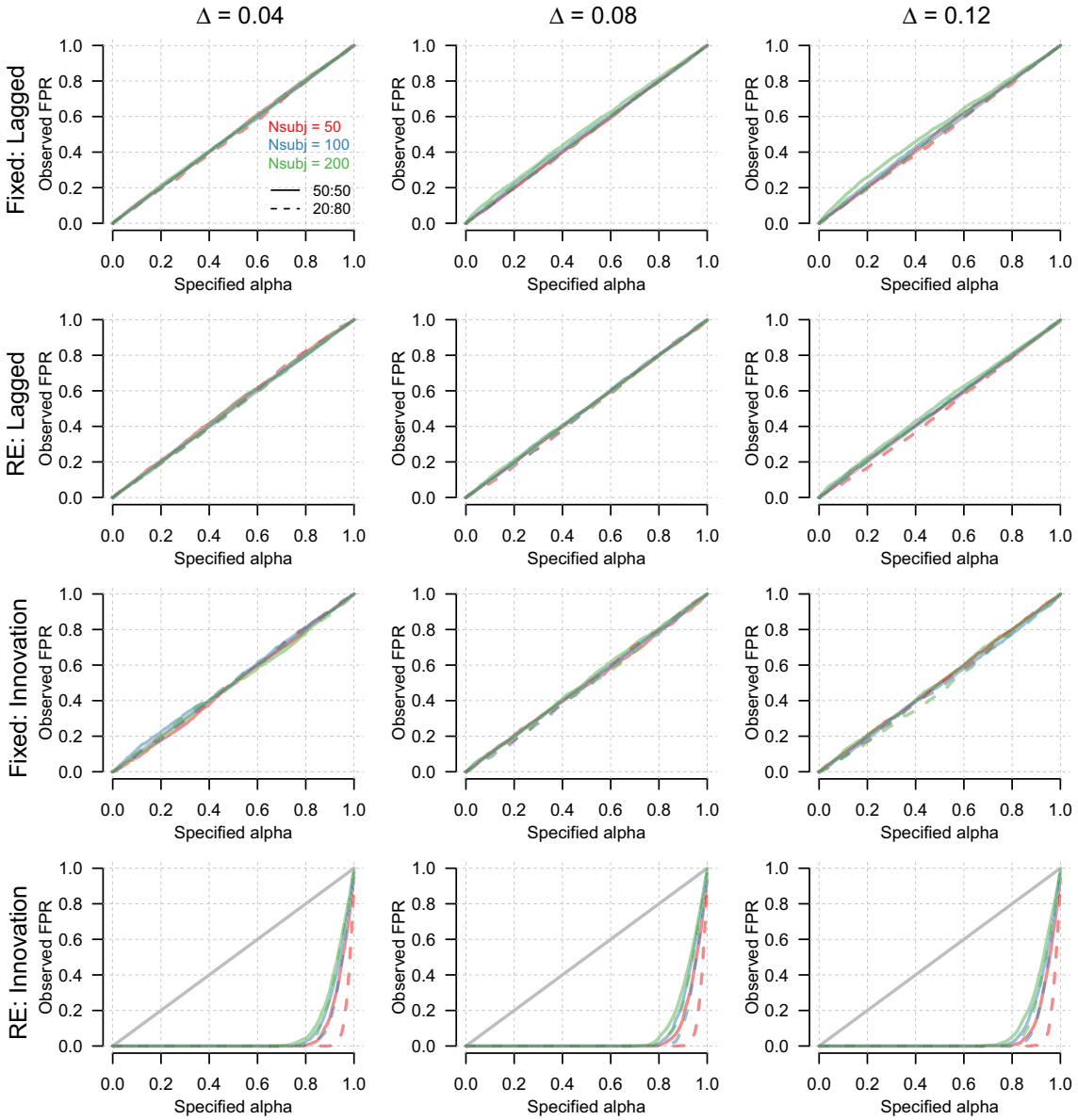


**Fig. 7** Specified $\alpha$-values ($x$-axis) plotted against the observed FPR in the simulation study, separately for all simulation conditions for the nonparametric permutation test

but still present. This suggests that it is unlikely that a single explanation can account for this phenomenon.

Second, we built the model up from scratch to see when the phenomenon would occur. We began with a standard (non-multilevel) VAR model and did not find the phenomenon there. We considered a multilevel model with only differences in fixed lagged effects, with all random effects variances being zero and innovations being uncorrelated. Here we found the effect. We built up the model in seven steps towards the model we have in our simulation, all showing the phenomenon in different magnitudes. This suggests that no simple explanation can be found. Our best hypothesis is that the reason for the increase in variance of sampling distributions and the resulting inflated FPRs is the misspecification inherent to the two-step mlVAR method we used in our paper to estimate multilevel VAR models.

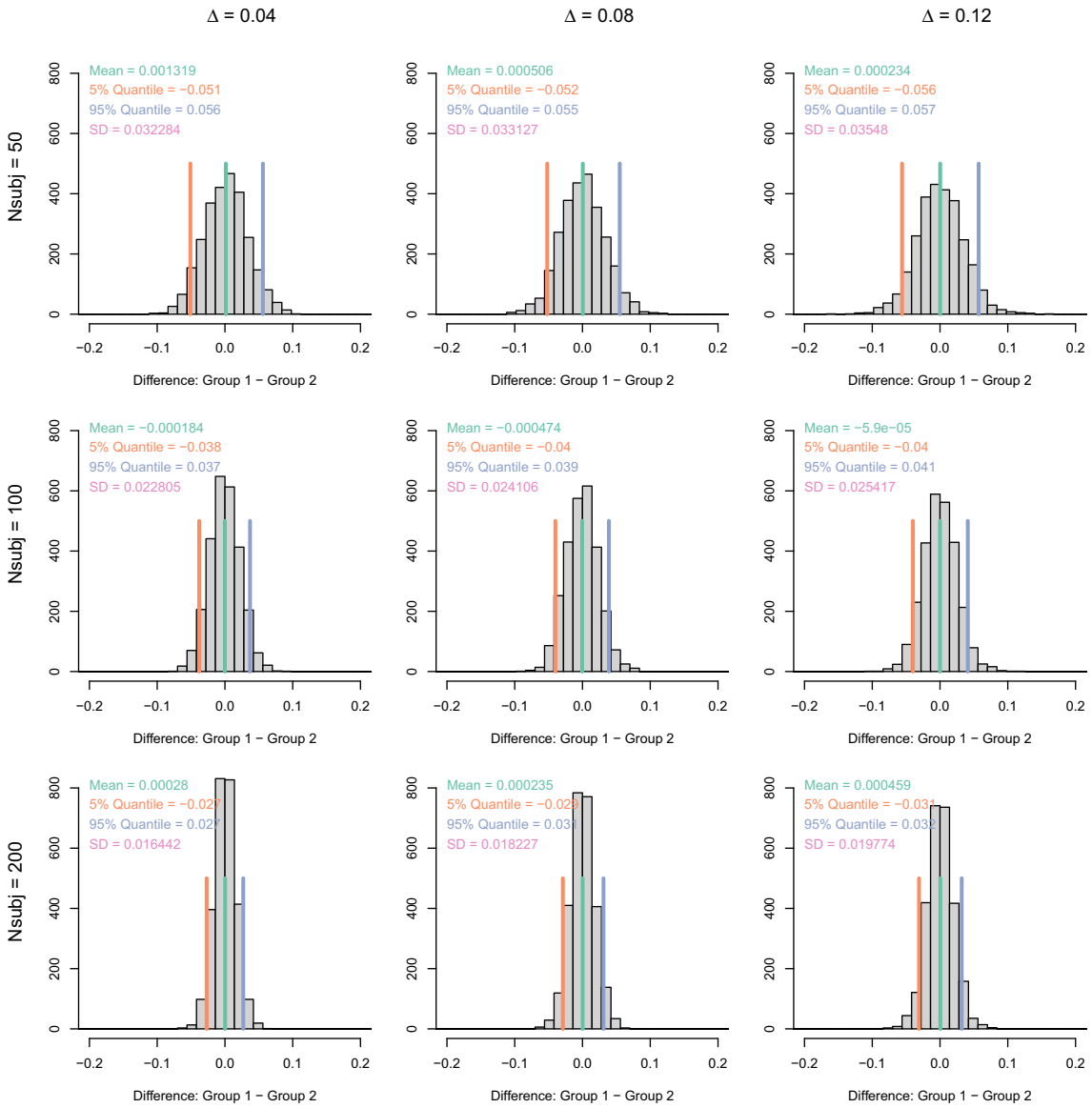


**Fig. 8** Sampling distributions of group differences for those parameters without group differences in the population model, separately for the different simulation conditions

## E. Estimating group differences in random effects variance in relations between innovations

The results for recovering group differences in random effects variances of relationships between innovations shown in Fig. 2 are at first sight puzzling: the three plotted specified $\alpha$ values 0.01, 0.05 and 0.1 correspond to a FPR of 0, but still the overall ROC curves reaches the top-left corner of the ROC space and therefore indicates high performance of the test. Figure 7 in Appendix D provides a more detailed picture of this misalignment, where we see that specified $\alpha$ across the whole range [0, 1] are highly misaligned with the observed FPRs. Here, we describe these results in a bit more detail and then explain them with a peculiarity in the estimation algorithm of the *mlVAR*-package we use to estimate multilevel VAR models.

Figure 9 shows the specified $\alpha$-levels ($x$-axis) plotted against the observed FPR (left) and the observed TPR (right) on the $y$-axis, fixed for the group difference $\Delta = 0.08$. We see that $\alpha$s until around 0.80 result in a FPR of 0 and for $\alpha > 0.8$ the FPR sharply increase towards 1. On the other hand, the TPR (or sensitivity) increases as a function of $\alpha$ exactly as one would expect: The higher $\alpha$, the easier it is for a group difference to be included. It also makes sense that the curves are shifted to the left if groups are balanced and larger, since in those situations group differences are easier to detect. However, it is odd that for $\alpha \in [0, 0.8]$ the TPR increases but the FPR remains constant at 0. In any reasonable test one would expect that that different $\alpha$s lead to different trade-offs

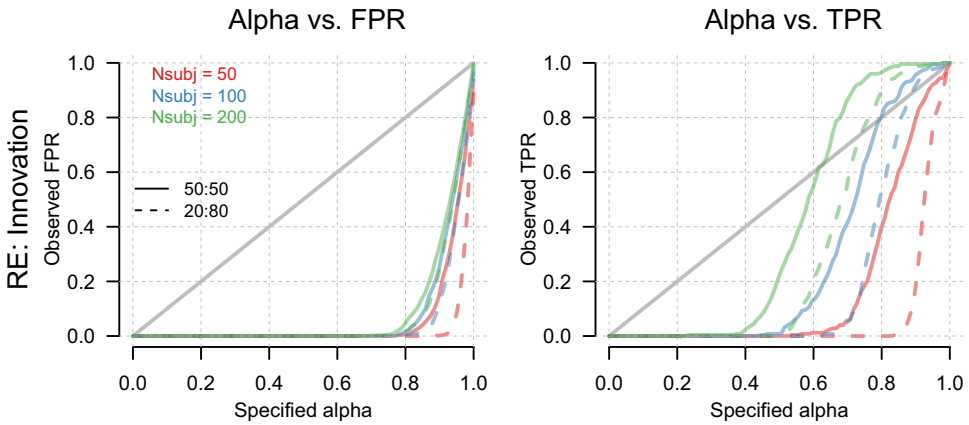


**Fig. 9** Specified $\alpha$-values ($x$-axis) plotted against the observed FPR in the simulation study, separately for all simulation conditions for the nonparametric permutation test

between FPRs and TPRs. That is, if we are increasing $\alpha$, we should get a higher TPR but also a higher FPR.

Explaining this result required investigating the precise estimates of the *mlVAR* package, which we use to estimate multilevel VAR models. Specifically, it turned out that random effects variances close to zero could often not be estimated by *mlVAR*, which itself uses the *lme4* package to estimate nodewise multilevel models. This means that the random effects variance that *were* estimated were biased away from zero, and therefore also group differences computed from these estimates were biased away from zero. This leads to sampling distributions under the null hypothesis with a variance that is strongly biased upwards.

This upwardly biased variance in the sampling distributions of the permutation test explains our findings in the following way. With standard $\alpha$s around 0.05 no group differences are detected at all, because the variance of the sampling distribution is too large. When increasing $\alpha$, one picks up true positive group differences first, since those are relatively large compared to false positive group differences. This explains the increase in TPR as a function of $\alpha$ in the right panel of Fig. 9. Only when increasing $\alpha$ beyond 0.80 it is high enough to also include false positive group difference, which are much smaller than the true positive ones. This explains the sharp increase in the FPR for $\alpha > 0.8$. This situation is far from desirable, and essentially renders the test for this parameter useless. In principle, if the simulation setup is very close to the data analytic situation at hand, one may pick $p$ values based on our simulation results to obtain a desired FPR. However, we do not know how this problem generalizes to situations beyond the ones considered in our simulation and therefore do not recommend this procedure in such situations.

**Acknowledgements** We would like to thank Inga Marie Freund for helpful feedback on earlier versions of this manuscript. We would like to thank Koval et al. (2013) for making their data openly available so we could use it in our tutorial.

**Funding** JMBH was supported by the project 'New Science of Mental Disorders' (www.nsmd.eu), supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.004.016). SE was supported by National University of Singapore grant A-8001098-00-00.

**Availability of data and materials** The data and code to reproduce the tutorial is available at https://github.com/jmbh/mlVARGD.

**Code Availability** The code to reproduce all simulations, analysis, and figures is available from https://github.com/jmbh/mlVARGD.

## Declarations

**Conflicts of Interest** None.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., ... Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8*(4), e60188.

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). Var (1) based models do not always outpredict ar (1) models in typical psychological applications. *Psychological Methods, 23*(4), 740.

Conner, T. S., & Barrett, L. F. (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. *Psychosomatic Medicine, 74*(4), 327–337.

Curtiss, J., Fulford, D., Hofmann, S. G., & Gershon, A. (2019). Network dynamics of positive and negative affect in bipolar disorder. *Journal of Affective Disorders, 249*, 270–277.

Dablander, F., Ryan, O., & Haslbeck, J. M. B. (2020). Choosing between ar (1) and var (1) models in typical psychological applications. *PLoS ONE, 15*(10), e0240730.

Elovainio, M., Kuula, L., Halonen, R., & Pesonen, A.-K. (2020). Dynamic fluctuations of emotional states in adolescents with delayed sleep phase–a longitudinal network modeling approach. *Journal of Affective Disorders, 276*, 467–475.

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.

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers. *Machine Learning, 31*(1), 1–38.

Good, P. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Hamaker, E. L. (2022). The curious case of the cross-sectional correlation. *Multivariate Behavioral Research,* 1–12.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science, 26*(1), 10–15.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton, NJ: Princeton University Press.

Haslbeck, J. M. B. (2020). Estimating group differences in network models using moderation analysis. *Behavior Research Methods,* 1–19.

Haslbeck, J. M. B. (2023). *Mnet: Modeling group differences and moderation effects in statistical network models*. R package version 0.1.0.

Haslbeck, J. M. B., & Epskamp, S. (2023a). Dependency of between person correlations on temporal effects. *PsyArXiv Preprint*. https://psyarxiv.com/e2qmx

Haslbeck, J. M. B., & Epskamp, S. (2023b). Observed correlations between person-means depend on within-person correlations.

Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2021). A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research, 56*(1), 120–149.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Kernel smoothing methods. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* 191–218.

Hoekstra, R. H., Epskamp, S., & Borsboom, D. (2022). Heterogeneity in individual network analysis: Reality or illusion? *Multivariate Behavioral Research,* 1–25.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., & Shmulevich, I. (2009). Fewer permutations, more accurate p-values. *Bioinformatics, 25*(12), i161–i168.

Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion, 13*(6), 1132.

Kuppens, P., Dejonckheere, E., Kalokerinos, E. K., & Koval, P. (2022). Some recommendations on the use of daily life methods in affective science. *Affective Science, 3*(2), 505–515.

Mansueto, A. C., Wiers, R. W., van Weert, J., Schouten, B. C., & Epskamp, S. (2022). Investigating the feasibility of idiographic network models. *Psychological methods*.

McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman: Hall/CRC.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in mplus. *Psychological Methods, 25*(5), 610.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*(3), 221–237.

Radloff, L. S. (1977). The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401.

Ryan, O., Dablander, F., & Haslbeck, J. M. B. (2023). Towards a generative model for emotion dynamics.

Snippe, E., Viechtbauer, W., Geschwind, N., Klippel, A., de Jonge, P., & Wichers, M. (2017). The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials. *Scientific Reports, 7*(1), 46523.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability, 57*(1).

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science, 23*(6), 466–470.

Vaart, A. v. d. (1998). *Asymptotic statistics*. New York: Cambridge University Press.

Van Borkulo, C. D., van Bork, R., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., ... Waldorp, L. J. (2022). Comparing network structures on three aspects: A permutation test. *Psychological Methods*.

Wood, S. N. (2017). *Generalized additive models: An introduction with r*. CRC press.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.