Testing for Group Differences in Multilevel Vector Autoregressive Models

Jonas Haslbeck*[1,2], Sacha Epskamp[3], and Lourens Waldorp[2]

[1]Department of Clinical Psychological Science, Maastricht University
[2]Department of Psychological Methods, University of Amsterdam
[3]Department of Psychology, National University of Singapore

Abstract

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

1 Introduction

Time series are becoming a common data type across disciplines in psychological research (Conner & Barrett, 2012; E. L. Hamaker & Wichers, 2017; Kuppens et al., 2022; Miller, 2012; Trull & Ebner-Priemer, 2014). 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 individual to estimate the VAR model (Bulteel et al., 2018; Dablander et al., 2020; Mansueto et al., 2022), which is why many researchers are estimating multilevel VAR models to time series data in psychology (Epskamp et al., 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 et al. (2020) investigated how multilevel VAR models fitted to self-reported emotion measures differed between individuals with and without sleep problems; Curtiss et al. (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 larger issue of distinguishing population inter-individual heterogeneity from sampling variation (Hoekstra et al., 2022).

In this paper, we address this issue by introducing two tests for group differences in multilevel VAR models. We lay out the rationale for a parametric test and a nonparametric permutation test and discuss their advantages and disadvantages based on theory; we then use a simulation study to determine how well group differences can be recovered with both tests in a set of scenarios mirroring applied research settings; to make the tests available to empirical researchers, we present the new R-package *mnet*, which implements both tests for the popular R-package to estimate multilevel graphical VAR models *mlVAR* (Epskamp et al., 2018). Finally, we provide a fully reproducible tutorial in which we test differences in emotion dynamics across groups with high vs. low depressive symptoms.

*jonashaslbeck@protonmail.com | www.jonashaslbeck.com

2 Parametric and Nonparametric Tests for Group Differences in Multilevel VAR Models

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

2.1 Multi-level 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 vector of means $\boldsymbol{\mu}$, and a linear combination of the same variables at the previous time point $\boldsymbol{B}\boldsymbol{y}_{t-1}$, and $\boldsymbol{\zeta}_t$ is a Gaussian innovation variance.

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 (Bulteel et al., 2018; Dablander et al., 2020; Mansueto et al., 2022). Many researchers therefore use multi-level VAR models, where parameters of each individual VAR model are shrunk towards the grand means of parameters across all subjects, which leads to estimates with less variance (Epskamp et al., 2018; McElreath, 2018; McNeish & Hamaker, 2020):

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

which has the same form as the standard VAR model above, except that all parameters now have an additional index for each subject s.

Two popular ways to estimate multi-level VAR models are currently the DSEM framework in MPLUS (Asparouhov et al., 2018), which fits the actual multi-level VAR model; and R-package *mlVAR* (Epskamp et al., 2018), which implements a 2-step approximation to the multi-level VAR model. The former is in principle preferred, however, it runs fast only with up to around $p = 6$ variables and is only implemented in the proprietary MPLUS software. The implementation in the *mlVAR* is fast since it estimates the model in several steps (for details of the algorithm see Supplement S1 in Epskamp et al. (2018)). This step-wise process has the downside that estimates are biased when the number of time points per subject is low (E. Hamaker, 2022; Haslbeck & Epskamp, 2023b; McNeish & Hamaker, 2020). However, since the *mlVAR*-method is the only one that is currently freely available, we will focus on this method in this paper.

The parameters that are typically of main interest in multi-level 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 parameters discussed above.

2.2 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. Since we do not know the population standard deviation of sampling distributions and often have groups with few subjects, we use Student's t distribution instead of the Gaussian distribution for the parametric test. Note that Student's distribution converges to the standard Gaussian distribution as the number of subjects goes to infinity.

2.3 Nonparametric Permutation Test

In a permutation test, we construct a sampling distribution under the null 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. We then estimate separate multilevel VAR models on these two permuted datasets and compute a difference for each parameter. For the multilevel VAR model, this includes the parameters of the between-subjects network, the fixed effects and the random effects variances of the temporal network, and the fixed effects and the random effects variances of the contemporaneous (residual / innovation) network. By repeating this process many times, we obtain an empirical estimate of the sampling distribution for each parameter under the null hypothesis that the group difference is zero. We can then compare the observed group differences 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.

2.4 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 are measured not with Visual Analog Slider scales which could practically considered as continuous measurements, but instead Likert scales with ordinal 3-7 response categories. 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.

3 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 if different sizes in multilevel VAR models can be recovered with the permutation test in scenarios that are typical in applied research.

3.1 Simulation Setup

3.1.1 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 $\Delta \in \{0.05, 0.15, 0.25\}$, the total number of subjects $N_{\text{subj}} \in \{50, 100, 200\}$ and the

proportion of subjects in the first group $p_{G1} \in \{0.2, 0.5\}$. 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. It would have also been desirable to vary additional 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 we had to limit the simulation study to the most important parameters.

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 the VAR models typically found in empirical data in studies using the Experience Sampling Methodology (ESM). ESM time series often capture positive and negative affect items, which show the following stable patterns in the $p \times p$ matrix ϕ 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 et al., 2023). We implemented this by setting all autocorrelations $\phi_{i,i} = 0.4$. 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 ϕ 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 residual 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 residual partial correlations, and also with standard deviations of $\sigma_{\text{RE}} = 0.05$.

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 residual 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 5 group differences for partial correlations between means and fixed and random effects variances of residual partial correlations, and a fixed number of 10 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. After generating the model of Group 1 and the group differences, we check whether the models of both groups are stable VAR models. In the rare cases where this is not the case, we repeat the generation of the true models until both are stable. 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. For each individual we sample $N_t = 100$ time points.

3.1.2 Estimating Group Differences

We estimate group differences using the R-package *mnet* (Haslbeck, 2023), which performs the permutation test described above using the R-package *mlVAR* (Epskamp et al., 2018). We choose to estimate mlVAR models with uncorrelated random effects, because the correlated random effects that can be obtained from the mlVAR output are rarely of interest and the running time would be much longer for correlated random effects. Note that Haslbeck and Epskamp (2023a) recently showed that the 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 (see also E. Hamaker, 2022). Consequently, we do not expect to be fully able to recover group differences in between-networks. However, to raise awareness of this bias and to see to what extent group differences might still be recoverable in ideal conditions, we also estimate and evaluate group differences of between-networks. In line with the analysis in the previous section, we choose $N_p = 1000$ permutations

to obtain accurate inferences but at the same time render the simulation study feasible. To obtain estimates of the population performance, we run the design with 3 (Δ) $\times$ 3 (N_{subj}) $\times$ 2 (p_{G1}) $\times$ 2 (p) = 36 cells for 50 repetitions.

3.1.3 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 presented results are conditional 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 of the permutation test this tuning parameter is the significance threshold α 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 α chosen in practice will lie between those extremes and depend on the desired trade-off between false positives and false negatives. The code to reproduce the simulation study and all results in the paper can be found at https://github.com/jmbh/mlVARGD.

3.2 Results

3.2.1 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 contemporaneous effects, and partial correlations between personwise means (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. Therefore, we 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 the group difference is small ($\Delta = 0.05$), having only $N_{\text{subj}} = 50$ is not enough to reliably recover group differences. On the other hand, we already achieve high performance with $N_{\text{subj}} \in \{100, 200\}$. For medium group differences ($\Delta = 0.15$) we see that the performance with $N_{\text{subj}} = 50$ becomes better, but is still far from desirable. The performance with $N_{\text{subj}} \in \{100, 200\}$ is even higher in this case. Finally, for large group differences ($\Delta = 0.25$), we again see that performance stacks up as a function of N_{subj}, however, now also $N_{\text{subj}} = 50$ is sufficient for a reasonable recovery performance. We see that unbalanced groups lead to lower performance, but the effect is relatively small across conditions. Very similar patterns of results hold for the random effects variance of lagged effects (second row), the residual partial correlations (third row), and the random effects variances of the residual partial correlations (fourth row).

While the ROC curves are constructed using 100 α-values $\in [0, 1]$ we display the commonly used α values $0.01, 0.05$ and 0.1. We see that they are located in the top left of the ROC-curve, reflecting the fact that these α-values typically reflect a reasonable trade-off between FPR and TPR. If the test works perfectly, these α-values should line up with the corresponding TPR. For example, for $\alpha = 0.05$ the FPR should be 0.05. We see that this is largely the case for all parameter types except the fixed lagged effects. In this case, the TPR of a fixed α seems to depend somewhat on the size of the group difference Δ.

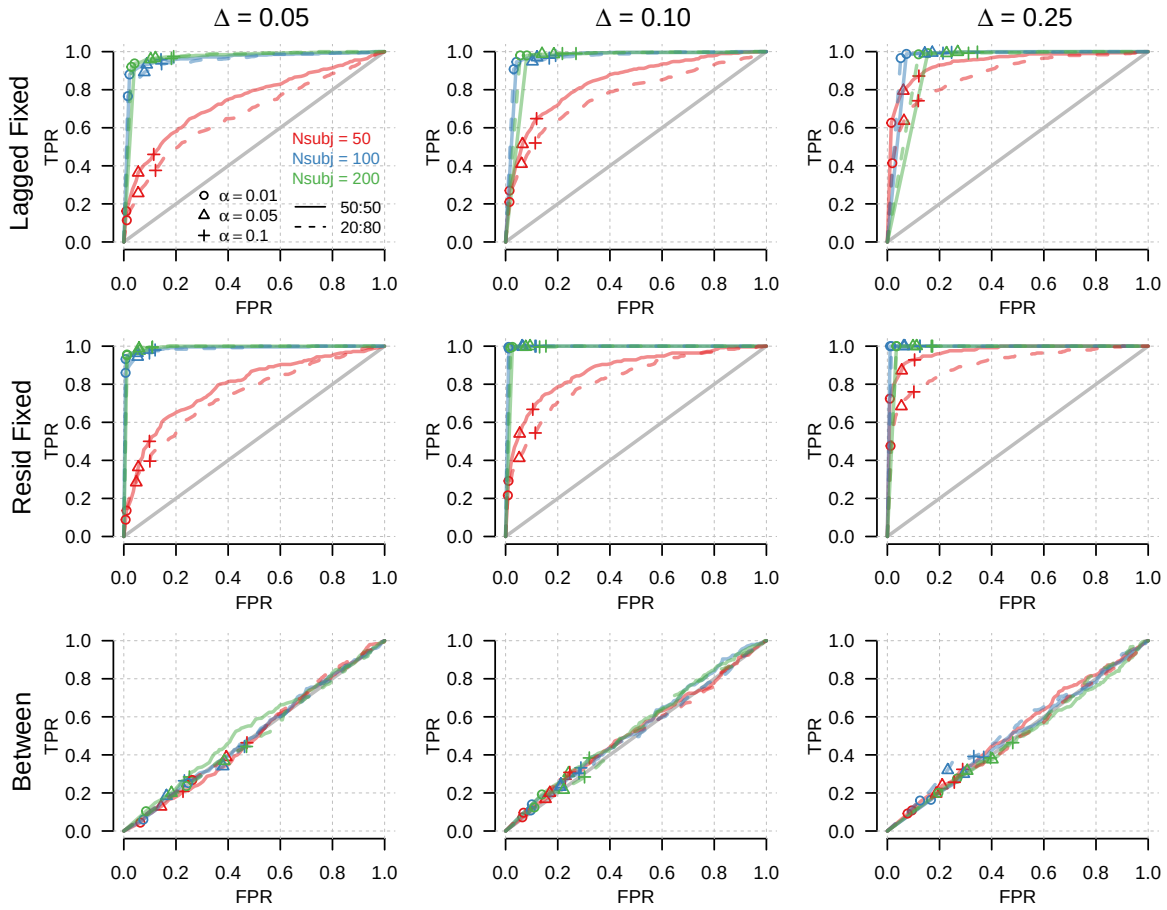

Figure 1: ROC curves of the parametric test for different numbers of total subjects (colors) and proportions in the first group (dashed vs. solid lines) for group differences of different sizes (columns) for the three variable types (rows) for which the parametric test could be performed. The diagonal lines indicate random performance. While ROC curves are constructed using 100 α-values from 0 to 100 we display the three popular choices $0.01, 0.05$ and 0.10.

In contrast to the first four parameter types, the performance for recovering group differences in networks capturing partial correlations between person-specific means (bottom row) is extremely low. As discussed above, the reason is that the 2-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 residual 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).

3.2.2 Results for Nonparametric Permutation Test

The results for the nonparametric permutation test are shown in Figure 2. The qualitative results are similar to those of the permutation test, except that we are also able to recover group differences in random effects variances of the lagged and the contemporaneous effects. For random effects of partial correlations between residuals/innovations, we see that the displayed α-values are all at FPR $= 0$ & TPR $= 0$. This is because all p-values for this parameter were very large, requiring much higher α to reject the null hypothesis. Note that we do not see results for the unbalanced group for the between-person network. This is because the permutation test is performed on partial correlations between personwise means, which are only defined when a random effects variance was estimated for these means. This was typically not the case for the group with only 12 subjects in the unbalanced $N_{subj} = 50$ condition. We again see that α values roughly correspond to the expected FPRs. The exception are again fixed lagged effects, in which the FPR rates are larger than expected, and this effect is larger for larger group differences Δ.

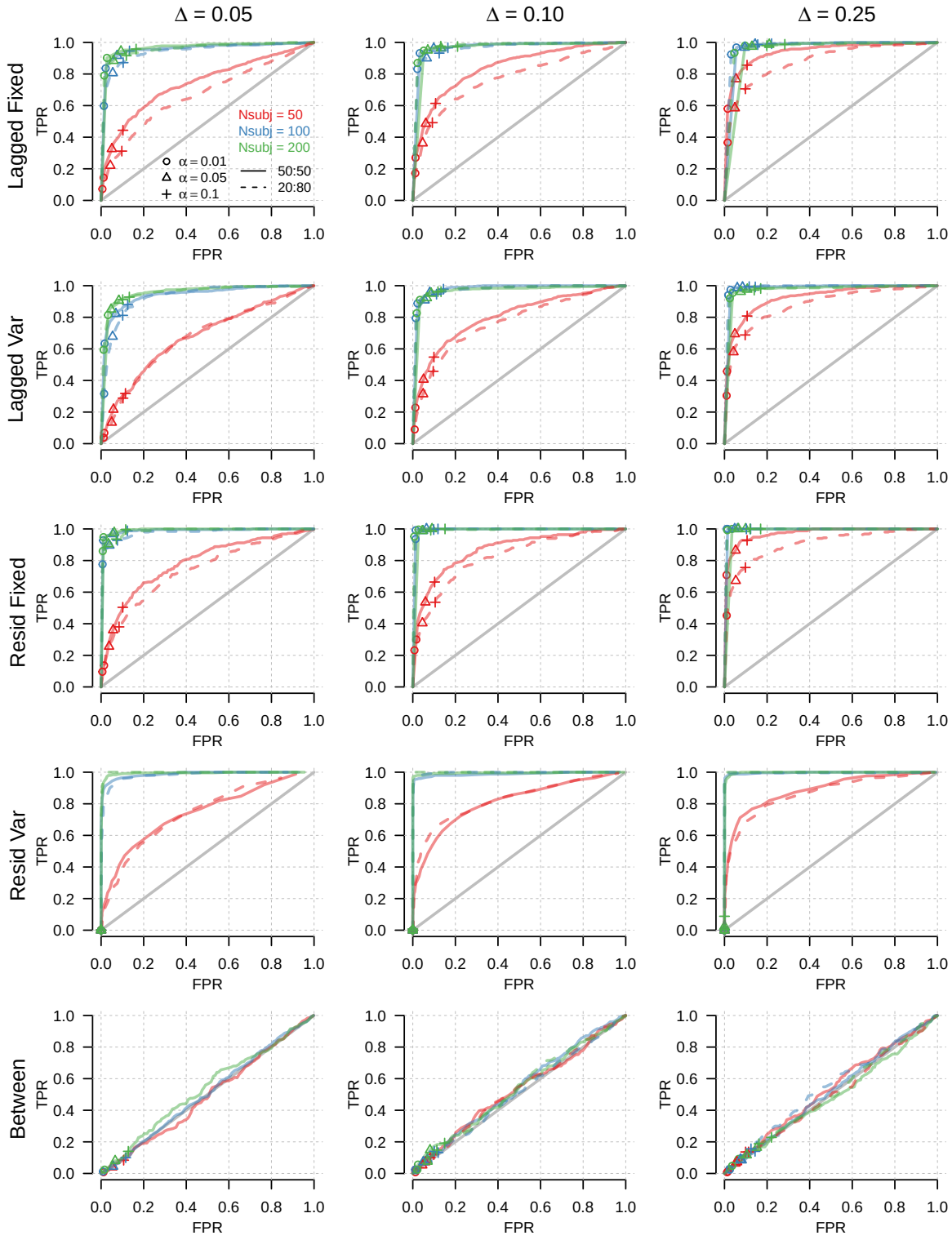

Figure 2: ROC curves for different numbers of total subjects (colors) and proportions in the first group (dashed vs. solid lines) for group differences of different sizes (columns) for the five considered variable types (rows). The diagonal indicates random performance.

3.2.3 Comparing Results of Both Tests

To better compare the performance of the parametric and nonparametric permutation tests, we compare their performance for the three parameter types in which the comparison is possible in Figure 3. Here we have averaged over balanced and unbalanced groups for each test and display the performance of the permutation test in solid lines, and the performance of the parametric test in dashed lines. We can see that the performance of the two tests is very similar.

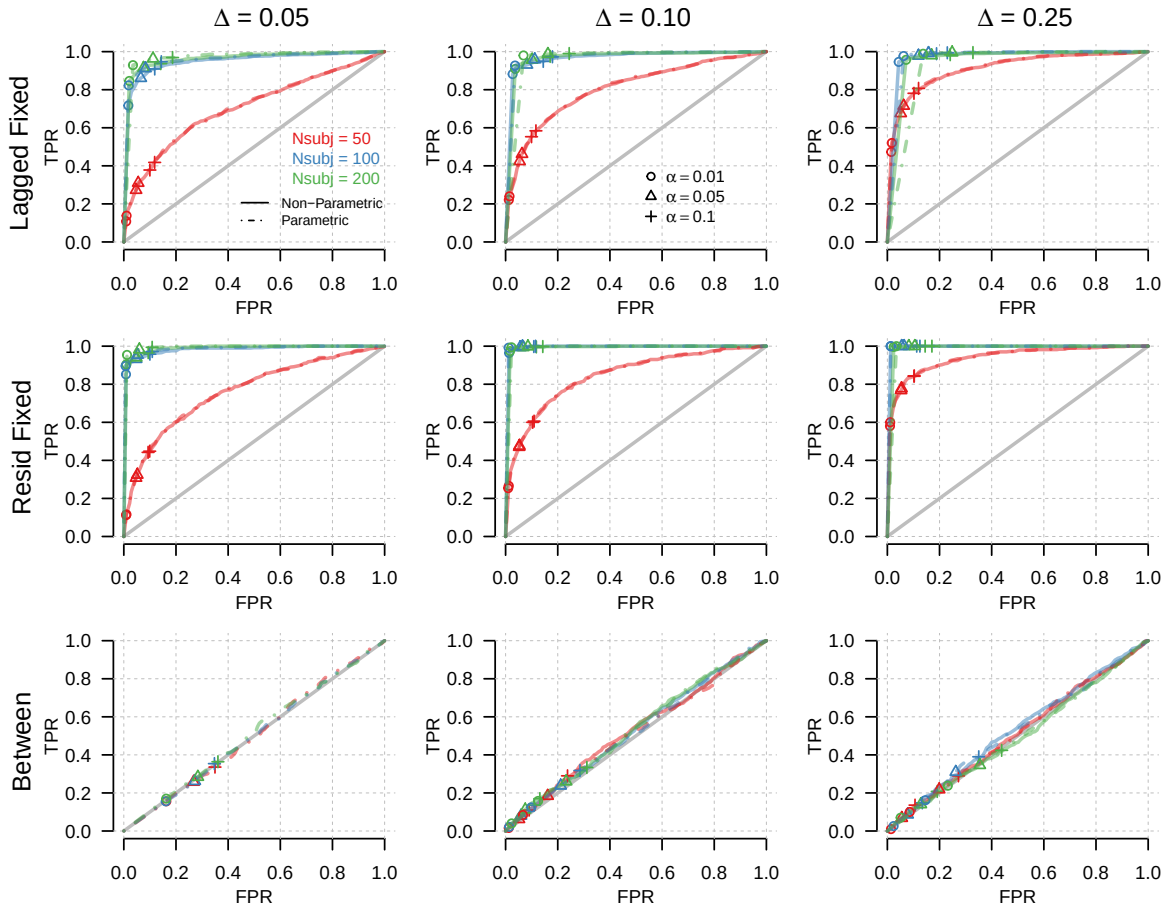

Figure 3: ROC curves for different numbers of total subjects (colors) and type of test (solid lines = permutation test; dashed lines = parametric test) for group differences of different sizes (columns) for the five considered variable types (rows). The diagonal indicates random performance.

3.2.4 Discussion of Results

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. We have seen that $N_{subj} = 50$ is insufficient to estimate group differences reliably if the true differences are small $\Delta = 0.05$ or medium sized $\Delta = 0.10$. However, already with $N_{subj} \in \{100, 200\}$ we obtain relatively high performance also for smaller group differences. Unbalanced groups had a strong negative impact on performance only for the condition with few $N_{subj} = 50$. We also saw that the performance was high for all parameter types except the "between-person" network, which we explained by a bias in the 2-step procedure used in *mlVAR* (Haslbeck & Epskamp, 2023a).

While α threshold values lined up with the expected FPR for most parameters, it is consistently too large for fixed lagged effects and appears to be larger for larger group differences Δ. This effect occurred for both the parametric and the nonparametric test, suggesting that it is not a specific problem of either test. The remaining explanation therefore has to be due to misspecification. The only possible source of misspecification is the 2-step estimation procedure implemented in *mlVAR*, since it does not exactly fit the data generating multilevel VAR model. We expect that when performing the two tests based on an estimator of the exact multilevel model such as in DSEM (Asparouhov et al., 2018), the α 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 residuals 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. Other reasons to preferring the permutation test are that it allows group differences to be tested also

for random effects variances. And finally, the permutation test allows group comparisons for within-subjects designs, for example to compare a group before and after an intervention. The downside of the permutation test is that it is computationally intensive.

3.2.5 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_{subj} and decreasing Δ leads to the same SNR and therefore the same results. In addition, did not correlate the random effects of lagged and residual 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.

3.3 R-Tutorial: Application to Emotion Time Series Data

We use data from the Experience Sampling Method (ESM) study of Koval et al. (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 which are scored on a 0-3 scale. We use the clinical cutoff of the CES-D ≤ 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 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 of 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 perturbation 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 minutes 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 Figure 4. The complete code to reproduce the figure from the output of `mlVAR_GC()` can be found at https://github.com/jmbh/mlVARGD.

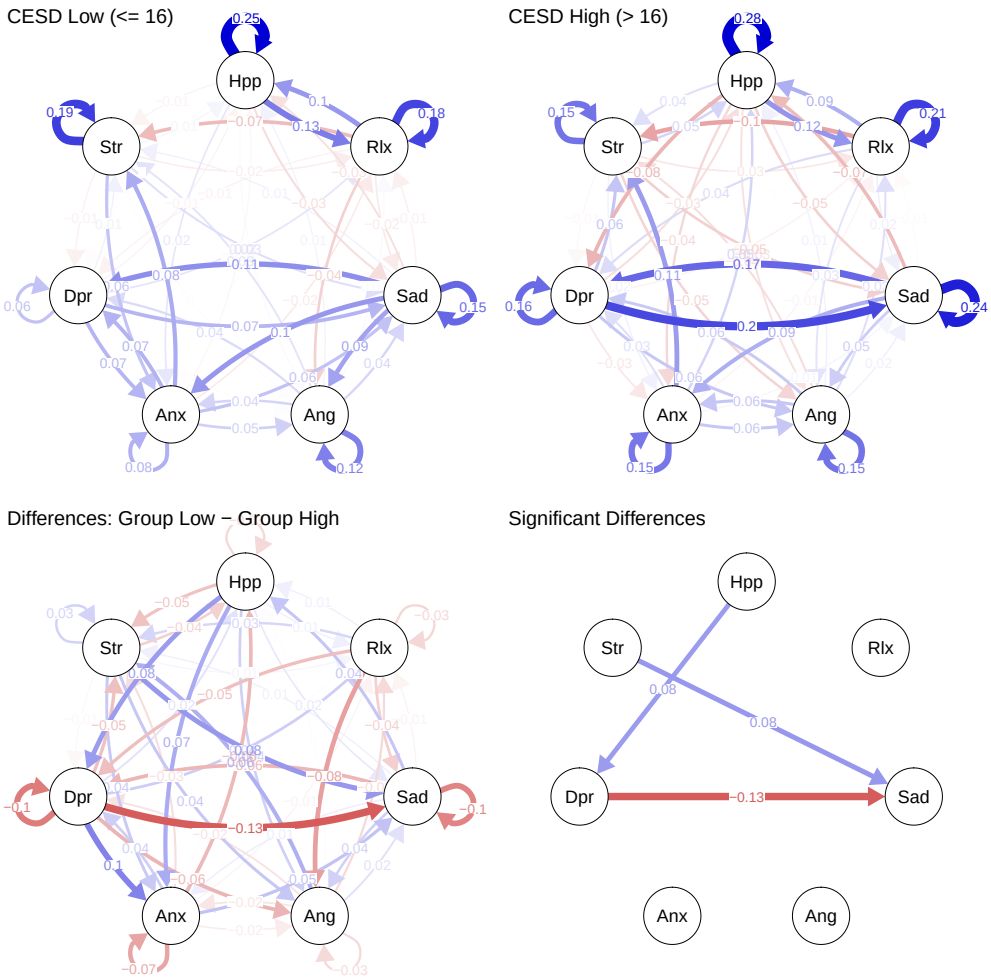

Figure 4: Inspecting group differences in fixed lagged effects in the data of Koval et al. (2013). Hpp = Happy, Rlx = Relaxed, Sad = Sad, Ang = Angry, Anx = Anxious, Dpr = Depressed, Str = Stressed; The top panel shows the fixed lagged effects of the two groups; the bottom left panel shows group differences (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)

The top panels of Figure 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 significant 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 (Bulteel et al., 2018; Dablander et al., 2020; Mansueto et al., 2022).

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 Figure 2 we see that, for threshold $\alpha = 0.05$, we expect a TPR of around 0.90 and a FPR of around 0.05.

4 Discussion

In this paper 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 of 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 small group differences ($\Delta = 0.05$) can already be recovered relatively reliably (TPR ≈ 0.8, FPR ≈ 0.05) with 100 subjects. We found that unbalanced groups ($20:80$ vs $50:50$) only considerably reduced performance in the condition with 50 subjects. 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.

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 residuals and between-person heterogeneity in empirical data. However, 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 residuals 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 presented tests allow researchers to make inferences about groups differences for the fixed effect (parametric test) and both the fixed and random effects variances (permutation test) of lagged effects and residual 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 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 MPlus (Asparouhov et al., 2018; McNeish & Hamaker, 2020).

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 to this end 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 CESD 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 essentially assuming that this functional form takes 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 et al., 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 estimate them accurately (see Appendix A). For example, Knijnenburg et al. (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.

Acknowledgements. We would like to thank Inga Marie Freund for helpful feedback on a 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. 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.

Materials. Code to reproduce all results is available from https://github.com/jmbh/mlVARGD.

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

A Choosing the Number of Permutations

The computational cost of estimating a multilevel VAR model is relatively high 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 $N_{subj} = 100$ subjects per group with the mlVAR package takes around 1.2 minutes on a 2.6 GHz Intel i7 core. Therefore, running $N_P = 1000$ permutations sequentially would already take more than 21 hours. 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:

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

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 $\text{sd}(\hat{P}) = 0.005$ acceptable and the true p-value P is equal to 0.05, then we need $k \geq \frac{P(1-P)}{\text{sd}(\hat{P})^2} = \frac{0.05(1-0.05)}{0.005^2} = 1900$ permutations. On the other hand, if the p-value is extremely small, only very few permutations are needed: $k \geq \frac{P(1-P)}{\text{sd}(\hat{P})^2} = \frac{0.001(1-0.001)}{0.005^2} \approx 40$. 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 cut-off 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 required number of permutation to obtain good estimates for population sampling distributions and p-values in Figure 5.

The top left panel of Figure 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 any other underlying distribution. 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 Equation 1, the small p-value associated with $\Delta_{\text{crit}} = 0.045$ converges extremely quickly. In contrast, the largest considered p-value associated with $\Delta_{\text{crit}} = 0.021$ 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-values 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 will specify in the following Section. However, as is also clear from Equation 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 cut-off 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.

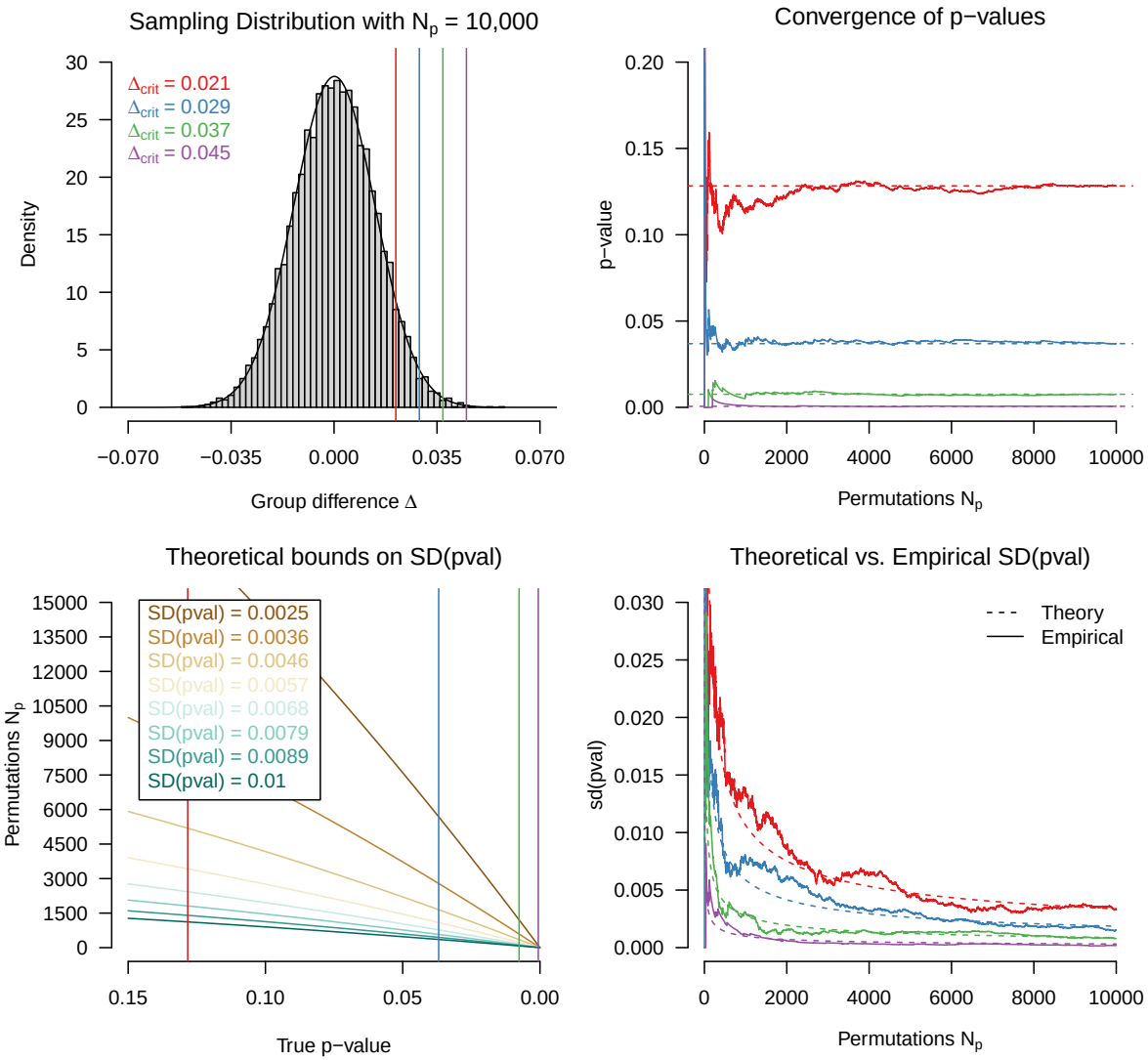

Figure 5: Top left: histogram of sampling distribution obtained with $N_p = 10,000$ permutations. We choose four critical values that are associated with p-values close to typical thresholds like 0.05 and 0.01; Top right: estimated $\hat{P}$ associated with the four critical values as a function of N_p;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