# Model Checking for Vector Autoregressive Models

Jonas Haslbeck*[1], Joran Jongerling[2], Björn Siepe[3], Sacha Epskamp[4], and Lourens Waldorp[1]

[1]Department of Psychological Methods, University of Amsterdam
[2]Department of Methodology, Tilburg University
[3]Department of Psychology, Philipps-Universität Marburg
[4]Department of Psychology, National University of Singapore

November 24, 2025

**Abstract**

Time series have become pervasive in psychological research and Vector Autoregressive (VAR) models have become one of the most popular classes of models to study within-person dynamics in such data. However, systematic checking of how well a VAR model fits the data is hardly ever performed. This is a problem, because model misfit can lead both to incorrect interpretations of model parameters and to missing effects in the data that would be theoretically interesting. We provide a tutorial that explains the theory behind model checking, introduces the most common types of VAR model misspecification in the context of psychological time series, and introduces diagnostics for them, using plots and simulations. We then apply these tools to assess model fit for a multilevel VAR model estimated on a typical empirical dataset of emotion measurements over three weeks of 179 persons. We conclude by discussing three complementary areas of research that could improve the modeling of psychological time series in the future.

## 1 Introduction

Psychological research is increasingly making use of intensive longitudinal data, facilitated by ubiquitous mobile devices, and collected with designs such as daily diary studies and Ecological Momentary Assessment (EMA) (Conner & Barrett, 2012; Fritz et al., 2024; Hamaker, 2025; Hamaker & Wichers, 2017; Kuppens et al., 2022; Miller, 2012; Trull & Ebner-Priemer, 2014). These data are sampled at relatively high frequency within the daily lives of participants and thereby allow researchers to study the within-person dynamics and how people differ in those dynamics in unprecedented detail. Capturing these dynamics in multivariate intensive longitudinal (or time series) data requires dedicated statistical models.

A natural starting point for studying the temporal dynamics in a time series is to directly model statistical relationships between variables across time points. The simplest multivariate model for such relationships is the Vector Autoregressive (VAR) model, in which each variable at a given time point is predicted by a linear function of all variables (including itself) at previous time points (e.g., Lütkepohl, 2005). Because estimating VAR models from the time series of a single person is often not feasible (Bulteel et al., 2018; Dablander et al., 2020; Mansueto et al., 2023) and to model heterogeneity across persons, many researchers use multilevel VAR models to analyze their time series data (Epskamp et al., 2018; McNeish & Hamaker, 2020). The fact that VAR models are the simplest multivariate time series models, together with the availability of software to estimate them, made multilevel VAR models one of the most widely used models for time series in psychological research (e.g., Blanchard et al., 2023; Haslbeck & Ryan, 2022; Peeters et al., n.d.; O. Ryan, Dablander, & Haslbeck, 2023).

The reporting on (multilevel) VAR models is largely focused on the fixed lagged-effects estimates, which are often visualized as directed networks (e.g., Veenman et al., 2024) and some studies also report associated random effects variances (e.g., Bringmann et al., 2013). However, hardly any studies perform systematic model checking for the VAR model (for a review of 43 published VAR papers, see Appendix A). That is, currently the common practice is not to perform any systematic analysis of residuals to assess overall model fit and to identify sources of misfit. This is despite the fact

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

that essentially every textbook on regression analysis recommends some form of model checking (e.g., Draper, 1998; Fox, 2015; Gelman et al., 2021; McCullagh, 2019; McElreath, 2018; Montgomery et al., 2021; T. P. Ryan, 2008) and VAR models consist of a number of linear regression models.

Model checking is important for two reasons. First, if a VAR model does not fit the data well and it is the only way in which we explore the data, one may miss systematic effects in the time series that are theoretically highly interesting. This is plausible because most researchers would agree that human functioning is more complicated than a VAR model, and consequently, the dynamics in the measurements provided by humans are unlikely to be fully captured by a VAR model (Haslbeck, Ryan, Robinaugh, et al., 2022). Second, checking the model fit is important to ensure that the estimated model parameters can be meaningfully interpreted. If a VAR model fits the data poorly, interpreting VAR parameters as if the model fit the data well could lead to grave misrepresentations of the actual effects in the data. For example, in the presence of an unmodeled trend, we might interpret the VAR parameters as describing a highly stable system, while the system is not stable at all (e.g., O. Ryan, Haslbeck, & Waldorp, 2023). This extends to the question of heterogeneity between persons: if there is considerable model misfit for a large proportion of persons, it would be misleading to present fixed effects estimates of multilevel VAR models as the average effects across persons.

A topic adjacent to model fit is how well a model predicts future observations. While predictive performance is a poor proxy for model fit, it is interesting in almost all applications and yet under-reported. For example, we would almost always like to know whether the predicted values are close to observed values at the next time point or whether the predictions are usually far off. While this information is crucial for anyone interpreting a VAR model, it is especially important when using time series models to inform preventive interventions (Dorais, 2024). Next to checks on model fit, we will therefore also show how to assess prediction errors for every variable and person, together with ideas for how to report them.

If residual analysis is so important for model construction and model interpretation, why do not more researchers apply it to multilevel VAR models? We think that the reasons include (a) that it is not always straightforward to obtain the random effects estimates from model objects and compute residuals; (b) that the amount of checking can seem overwhelming for the typical time series data comprised of multivariate time series of many persons; and (b) a decent amount of coding skills are needed to compute residuals and visualize them in a convenient way. Taken together, this has led to the common practice of not performing model checking on VAR models, with the consequence that it is often unclear to what extent the conclusions drawn from published VAR models are actually supported by the data.

In this paper, we tackle this issue by providing an accessible and fully reproducible tutorial on how to perform model checking for (multilevel) VAR models. We begin by introducing the basic theory behind model checking and two types of diagnostic tools to spot model misfit: The first type uses standard residual analysis comparing observations and predictions and evaluating whether residuals are normally distributed and do not change systematically across time. The second approach simulates new time series from the estimated models and compares them to the empirical time series akin to Bayesian Posterior Predictive Checks (Berkhof et al., 2000). We then consider a number of simulated examples with the most common forms of VAR model misspecification and discuss how to spot them with the different diagnostic tools. We then apply these tools in a typical empirical application in which we perform model checking on a multilevel VAR model estimated on a typical EMA time series of emotion measurements. We discuss how the detection of model misfit is not only a safeguard against interpreting parameters incorrectly, but can also be used to guide model extensions that can lead to the discovery of additional theoretically relevant dynamics. We end with discussing the role of good measurement in relation to model fit, various model extensions to address common types of misfit, and the connection between time series modeling and building theories.

## 2 Model Checking: Theory and Diagnostics

In this section, we introduce different strategies to assess model fit. We first sketch the basic theory of models, misspecification, and residual decomposition using the AR(1) model (Section 2.1). We present a number of standard plots to diagnose the extent and nature of misspecification. These plots either show residuals in different ways (Section 2.2.1), or compare data simulated from the estimated model with empirical data (Section 2.2.2). In Section 3, we then apply these tools to a number of simulated time series showing different scenarios of perfect model fit, and the most common forms of model misfit. With these intuitions, we will then diagnose model misfit in a typical empirical application of a multilevel VAR(1) model estimated from a time series of emotion measurements (Section 4).

## 2.1 Theory: Model, Misspecification, and Residuals

Here we provide a standard introduction to time series models, model misspecification, and residuals. This defines the problem and motivates the diagnostics we discussed below. We introduce the residual with the simpler lag-1 autoregressive, or AR(1), model, but everything we discuss also extends to (multilevel) VAR models.

We consider the following model:

$$Y_t = \theta + \phi Y_{t-1} + \varepsilon_t + m_t, \tag{1}$$

where $\theta$ is an intercept, $\phi$ is the autoregressive parameter, and $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$ are independent Gaussian innovations, centered at zero with a variance $\sigma^2$.

If the model consisted only of the intercept, the autoregressive part, and the innovations, then it would be an AR(1) model. However, the model contains an additional term, $m_t$, which represents model misspecification. This term can represent *any* deviation from an AR(1) model, which means that the model (1) in principle captures *any* time series model. Simple and common examples of misspecification are trends, non-linear effects, or changes in the distribution of innovations across time (also called heteroscedasticity). In the case of a simple linear trend, we could have $m_t = \beta t$. However, in general, $m_t$ can be a combination of many complicated terms.

If $m_t \neq 0$, we do not model all systematic effects in the time series with the AR(1) model. In this case, we can say that there is model misfit, or equivalently, that the AR(1) model is misspecified for the present data (generated by an unknown process in which $m_t \neq 0$).

The focus of this paper is to compare the predictions of an AR(1) (or VAR(1)) model to the observed data by inspecting their difference (i.e., the residuals) to determine whether $m_t \neq 0$, and if so, to characterize $m_t$ as well as possible. Finding that there is model misfit provides us with information about unmodeled systematic effects (i.e., $m_t$) that are theoretically interesting and give us ideas for how to extend the original AR(1) model. And it tells us to be less confident about our conclusions based on the AR(1) model if we choose to interpret it despite model misfit.

The comparison between predictions and observations is usually done by inspecting their differences, which are called residuals. Here we briefly explain how to obtain these residuals and how they help us to diagnose $m_t$.

We have the predictions of the AR(1) model:

$$\hat{Y}_t = \hat{\theta} + \hat{\phi} Y_{t-1}, \tag{2}$$

where $\hat{\theta}$ and $\hat{\phi}$ are estimated from the observed data generated from model (1). The residual is the difference between the observed value and the prediction, so:

$$e_t = Y_t - \hat{Y}_t = Y_t - (\hat{\theta} + \hat{\phi} Y_{t-1}). \tag{3}$$

Substituting the true data generating model (1) into the residual and rearranging yields:

$$\begin{aligned} e_t &= (\theta + \phi Y_{t-1} + \varepsilon_t + m_t) - \hat{\theta} - \hat{\phi} Y_{t-1} && (4)\\ &= \left[(\theta - \hat{\theta}) + (\phi - \hat{\phi}) Y_{t-1}\right] + \varepsilon_t + m_t. && (5) \end{aligned}$$

This shows that the residual $e_t$ has three components:

1. **Estimation error:** $(\theta - \hat{\theta}) + (\phi - \hat{\phi}) Y_{t-1}$,
2. **Misspecification error:** $m_t$,
3. **Irreducible error (innovations):** $\varepsilon_t$,

The decomposition provides us two important insights. First, if the estimation error is dominating the residual, then it will be difficult to diagnose whether the misspecification error $m_t \neq 0$. This would be the case when estimating an AR(1) model on a sample size that is too low. This is especially an issue for VAR(1) models, which have many more parameters, and therefore also require more observations to be estimated with low estimation error.

The second insight is that, if estimation error is negligible, we can inspect the distribution of $e_t$ and see whether it is different from the assumed distribution of $\varepsilon_t$. The specific way in which they differ will tell us about the unmodeled structure in $m_t$. Different types of unmodeled structures in the residuals can be detected with different diagnostics, which we introduce next.

Focusing only on the estimation of the AR parameters, the first two errors are mapping on the variance and the bias of an estimator for AR parameters (e.g., Chapter 6, Wasserman, 2004). For unbiased estimators, the estimation error takes the form of random variation around the true parameter; this error becomes negligible when the sample size increases. The misspecification error leads to bias on the AR parameters; this error remains no matter how large the sample size is. This means that model misspecification cannot be solved by collecting more data.

## 2.2 Diagnostics for Model Misfit

Here we introduce the two types of tools we will use to diagnose model misfit (see Figure 1), before using them to detect different types of model misfit in Section 3.

Before fitting any model, we recommend creating data visualizations such as line plots for every variable and every person, bivariate plots of relationships at the same time point or over lags of time, and histograms of distributions collapsed over time. We do not show these plots here, since we are focusing on model checking and because some of these plots are contained in the diagnostic plots we show.

### 2.2.1 Inspecting Diagnostic Plots

The first type of diagnostic inspects the residuals, which we obtain by subtracting the predictions from the empirical values (see Figure 1). The theory above showed that if there is no misspecification ($m_t = 0$), and estimation error is negligible, the residuals are equal to the innovations. We make clear assumptions about the distribution of these innovations: they are normally (Gaussian) distributed with a fixed mean $\mu = 0$ and variance $\sigma^2$ and independent across time. If the case of model misfit ($m_t \neq 0$), the residual is not equal to the innovations, and those assumptions are violated. That is, we might see residuals that are not normally distributed, that their variance changes across time, or that they are correlated across time points (e.g. Chapter 5, Hyndman & Athanasopoulos, 2021). The diagnostic plots we will consider in Section 3 allow us to test these assumptions. For example, showing a histogram of the residuals allows us to gauge whether they are Gaussian distributed, and plotting them across time allows us to see whether the distribution changes across time.

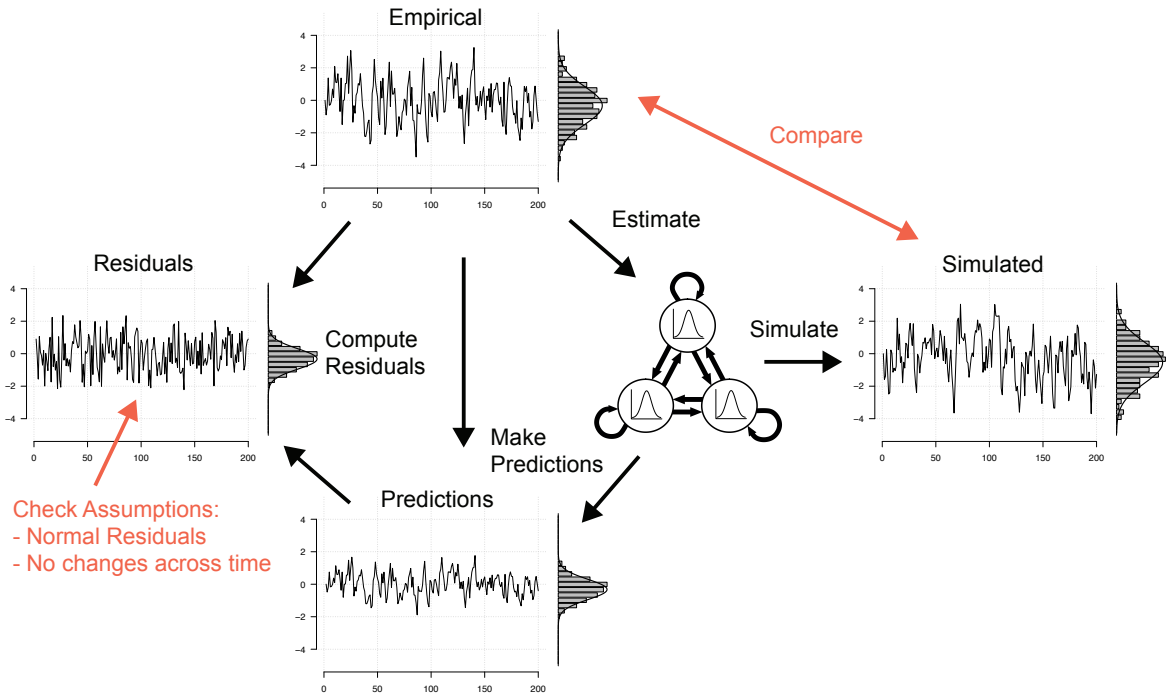


Figure 1: The two types of model checks discussed in this paper: First, computing the difference between the empirical data and the $p$-time step forward predictions of the VAR model estimated with the same data and checking whether the residuals are normally distributed and their distribution does not change over time; Second, simulating data from the estimated VAR model and comparing the simulated data with the empirical data.

### 2.2.2 Comparing Empirical Data with Data Simulated from Estimated Model

Another way to check model fit is to take the parameters of the estimated AR/VAR model, generate data from it, and see whether (summaries of) the generated data are similar to the corresponding (summaries of) the empirical data (see Figure 1). In a Frequentist setting, this could be called a parametric bootstrap diagnostic. In the Bayesian setting, this type of model checking is much more common and called a Posterior Predictive Check (PPC; Berkhof et al., 2000). In the Bayesian setting one draws repeatedly parameters from the posterior and simulates data from each draw, which allows one to assess the impact of uncertainty about parameter estimates on the simulated time series. This is a powerful tool that is able to spot types of misspecification that would be easy to miss with the diagnostic plots discussed above, as we will see in Section 3.

Simulating data from the fitted model can also be used beyond visually comparing the data to its empirical counterpart. We can compute any statistic on the time series and compare it between the empirical and the simulated time series. For example, one might be interested in the Root Mean Square Successive Difference (RMSSD) in the time series capturing to what extent the time series is changing from one time point to the next, which is likely difficult to eyeball. We can compute the RMSSD on the empirical time series and the simulated data and see whether they are similar. We can also simulate many time series and obtain a sampling distribution of RMSSD under the assumption that the fitted model is the true model. This allows us to make judgements that are not dependent on the sampling variation of a single simulation, thus enabling us to conduct Frequentist significance tests with the chosen statistic as the test statistic and the null hypothesis that the fitted model is true. This approach is a key strategy for model checking suggested in modern textbooks on regression (e.g., Gelman et al., 2021).

### 2.2.3 Limits of Diagnostic Plots & Formal Model Selection

Of course, model checking based on diagnostic plots has limitations. One is that some types of misspecifications cannot be easily spotted, especially if they are relatively small or if time series are very short. For example, it might be difficult to visually spot a small seasonal weekend effect in time series of only 2-3 weeks with typical innovation variance. Spotting such types of misspecification would likely require creating summaries over all persons of the dataset, such as plotting the residuals for different weekdays, across all persons (e.g. see Figure 4 in Siepe, Rieble, et al., 2025).

Next to diagnostic plots, misspecification can be detected with formal model comparison. For example, one could fit an alternative model with a weekend effect and perform model selection between the original and the extended model, using information criteria or prediction in a test-sample (see M. M. Haqiqatkhah & Hamaker, 2025, for a tutorial on seasonal models). While this approach is generally superior to laboriously spotting weekend effects in diagnostic plots, this modeling approach is only possible if we already know what we should be looking for. Finding that out is one of the main uses of the diagnostics we discuss in this paper.

# 3 Diagnosing Model Misfit in Simulated Examples

Diagnosing model misfit in empirical data can be difficult, because a model can be misspecified in multiple ways at the same time. We therefore first show how to use model checking diagnostics to spot one specific type of model misspecification at a time. To this end, we simulate data from models that are either correctly specified or misspecified in one of several particular ways. Equipped with these intuitions, we will then perform the same diagnostics to detect model misfit of a VAR(1) model fitted to empirical emotion measurements in Section 4. To keep the examples as simple as possible, we still stick to the simpler AR(1) model with which the model check diagnostics can be equally well explained. We provide the specification of all models discussed in this section in Appendix B.

In all examples we make use of data on a continuous scale, but the model checks we discuss can in principle also be used for Likert data. However, in the discussion we will also consider alternatives to the standard VAR model with Gaussian innovations which avoid this misspecification of the measurement domain.

In what follows, we use the above-discussed diagnostics to inspect the time series from two models that are correctly specified by the AR(1) model and therefore have perfect model fit; and three models for which the AR(1) model is misspecified. We discuss the three types of misspecification we think are most common in psychological time series, but discuss three additional ones in Appendix D.

## 3.1 Two Examples of Perfect Model Fit

We first consider two examples for which the AR(1) model fits perfectly to build intuition for how the diagnostic plots look in these cases. While both time series are generated by an AR(1) model, and therefore the AR(1) model has perfect model fit in both cases, in the first time series time points are dependent across time, while in the second time series they are *in*dependent. Comparing the model fit of those two models will show that predictability is a poor proxy for model fit.

*Example 1: AR Process with Dependence.* We begin with the time series shown in the first row of Figure 2. The first plot shows the empirical time series (black line) and the distribution resulting from collapsing all observations across time, shown as a histogram. The time series looks like it is varying around a stable mean, which is a type of dynamic that can be well approximated by an AR(1) model. Looking at the histogram of the distribution of values collapsed across time, we do not see large deviations from a Gaussian shape in the distribution, which could hint at non-Gaussian innovations.

Comparing the observations (black) with the predictions (orange) in the same plot, we see that the predictions seem to align well with the observed data, suggesting that the model predicts well one time step ahead. We can also quantify this predictive performance, for example, with the proportion of explained variance $R^2 = 0.29$ or a Root Mean Squared Error (RMSE), which here is 1.07. These are the prediction errors we would expect given the model we generated the data from (see Appendix C). In empirical data, the RMSE can be interpreted with respect to the response scale. For example, on a $[0, 100]$ scale, an RMSE of 5 would be relatively small, and 40 a very large prediction error.

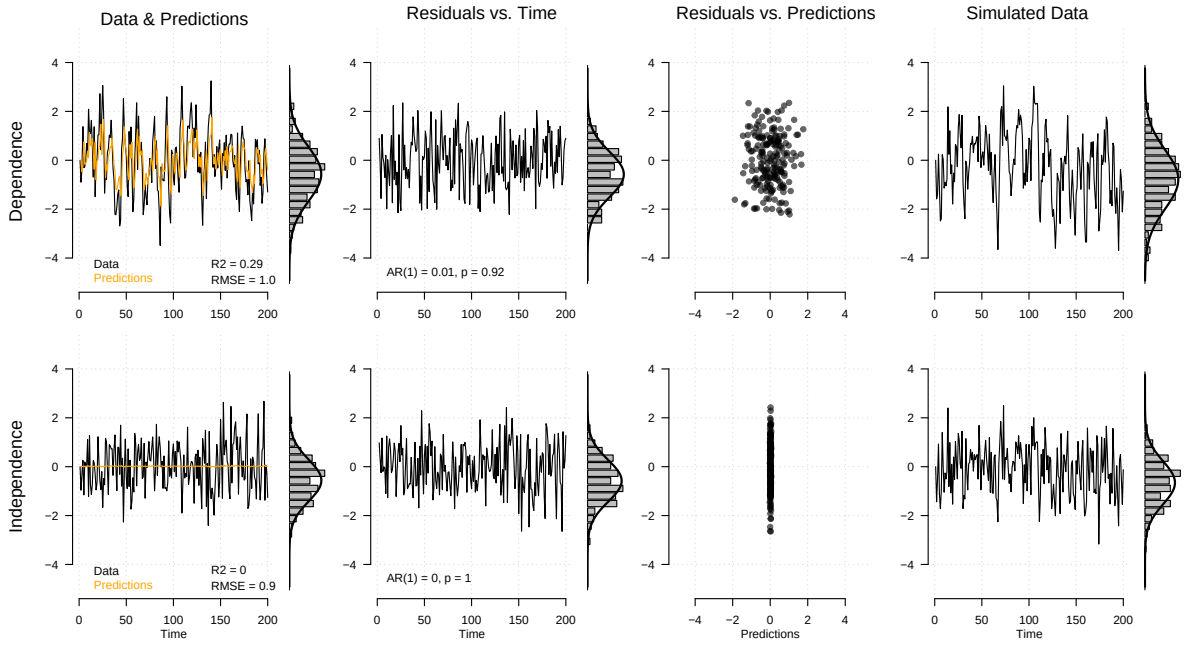


Figure 2: Two examples of time series for which the AR(1) model is correctly specified. In the time series on top, the observations are dependent which allows prediction. In the bottom time series, the observations are independent and out-predicting the mean is impossible.

The second plot showing the residuals across time suggests that the distribution of the residuals (e.g., mean and variance) stays the same across the entire time series. We computed the autocorrelation on the residuals which is equal to 0.01 ($p = 0.92$ under $H_0$: autocorrelation is zero), from which we conclude there is no evidence for unmodeled autocorrelation in the residuals. The histogram in the second plot shows the distribution of residuals, which is assumed to be Gaussian. Comparing the histogram to the density of the best-fitting Gaussian (black line) we conclude that the residuals are indeed approximately Gaussian. The third plot displays the residuals (y-axis) plotted against the predicted values (x-axis), from which we see that the distribution of the residuals is the same for all predicted values, suggesting that we did not miss any major non-linear effects in the model.

The last plot shows a single time series generated from the estimated AR model. We see that it looks similar to the empirical time series. It seems to vary around a similar mean, has the same variance, a similar autocorrelation, and a similar (Gaussian) distribution across time. Since the simulated data show the same characteristics as the observed data, we conclude that there are no large systematic effects in the data we did not model, and consequently, the model fits well.

*Example 2: AR Process without Dependence.* We next consider the time series in the second row of Figure 2. In the first plot, we see that the observed time series (black line) also seems to vary around a stable mean and that the distribution across time looks approximately Gaussian, suggesting that the AR(1) model fits the data well. Looking at the predictions (orange), we see that they are very close to zero at all time points. This is the case because the estimate of the autocorrelation coefficient is very close to its true value of zero, and we therefore predict with the mean, which here is equal to zero. We can also compute prediction errors, such as the proportion of explained variance, which is $R^2 = 0$ and the RMSE = 1, meaning that the model does not predict better than the mean (predicting 0 at all time points is the same as predicting the mean, which is 0).

The second plot shows residuals very similar to the first time series: the residuals are uncorrelated and have the same distribution across time, and the distribution is a Gaussian distribution centered at zero, as assumed by the AR(1) model. The third plot does not give us additional information above the previous plot for this time series since all predictions are roughly zero. The final plot showing the time series simulated from the estimated model looks similar to the empirical time series, suggesting that we did not miss any unmodeled structure and that the model is correctly specified for these data.

The analysis of the model fit of this second time series shows that model fit does not imply predictive performance. The second time series was generated from an AR(1) model in which the autocorrelation $\phi = 0$, meaning that all observations are independent across time (see Appendix B). Clearly, in such a situation we cannot hope to predict the next time step better than the overall mean. At the same time, the model checks show no model misfit, and since we generated the data from an AR(1) model, we know that the AR(1) model fits these data perfectly. This illustrates an important general point: low predictive performance does not indicate misspecification. Prediction tells us how well we can forecast, while model checking tells us whether the model has captured all systematic effects in the data.

## 3.2 Three Common Examples of Model Misfit

Next we consider three examples with common types of model misspecification. The definitions of the models we used to generate the example time series in this section are not needed for understanding the model checks we present here. Indeed, when analyzing empirical data, the data generating models are always unknown. However, we for the interested reader we provide the model definitions in Appendix B.

*Example 3: Switching between Stable States.* The first example time series, shown in the first row of Figure 3, seems to switch between two different "states": it either varies around $-2$ or around 2 and sometimes switches. When collapsing the data across time, we see that this behavior gives rise to a bimodal distribution, a phenomenon that seems to occur quite often in empirical data (Haslbeck et al., 2023). Looking at the predictions (orange) we see that they follow the data very closely, and a proportion of explained variance of $R^2 = 0.79$ and a rounded RMSE of 0.8 confirm that the AR(1) model predicts well one time step forward. This time series is therefore another example of how prediction performance is a poor indicator for model fit.

The plot showing residuals across time shows roughly Gaussian distributed residuals throughout the time series with the exception of extreme values at the time points where the switches occur. This is because the estimated AR(1) model has a very high $\phi$-coefficient (here $\hat{\phi} \approx 0.89$), which approximates the time series well as long as the next time point is in the same state. This is not the case when the system switches, explaining the extreme residuals with values $-4$ and 4. In addition to these extreme values, we also see a negative autocorrelation in the residuals, which also indicates model misfit.

The third plot showing the residuals as a function of the predicted values shows a clearly non-linear shape. We see that predictions cluster around $-2$ and 2 without any predictions in between and we see the four extreme residuals at the switching points. We also see that within the two groups of points, the residuals seem to be positively correlated with the predicted values. This plot clearly shows that the AR(1) model is misspecified for these data, which is in line with what we already know about the misspecification from the previous plot.

The time series simulated from the estimated AR(1) model, shown in the final plot, does show a highly autocorrelated time series, but does not exhibit the type of switching behavior we see in the observed time series. This suggests that we did not model all of the structure in the time series adequately. Of course, theoretical knowledge about the functional form and the dynamics allowed by an AR/VAR model would have also allowed us to conclude that such a model would be misspecified for this time series. The simulation approach to model checking, however, can also always be used without much theoretical knowledge. In practice, one would also simulate a number of time series. Here we only showed one for simplicity.

In this first example time series, the model misspecification was relatively easy to spot since we were able to spot it at least from three out of the four plots. However, some misspecifications do only show up in some of the diagnostic plots, as we will see in the next example.

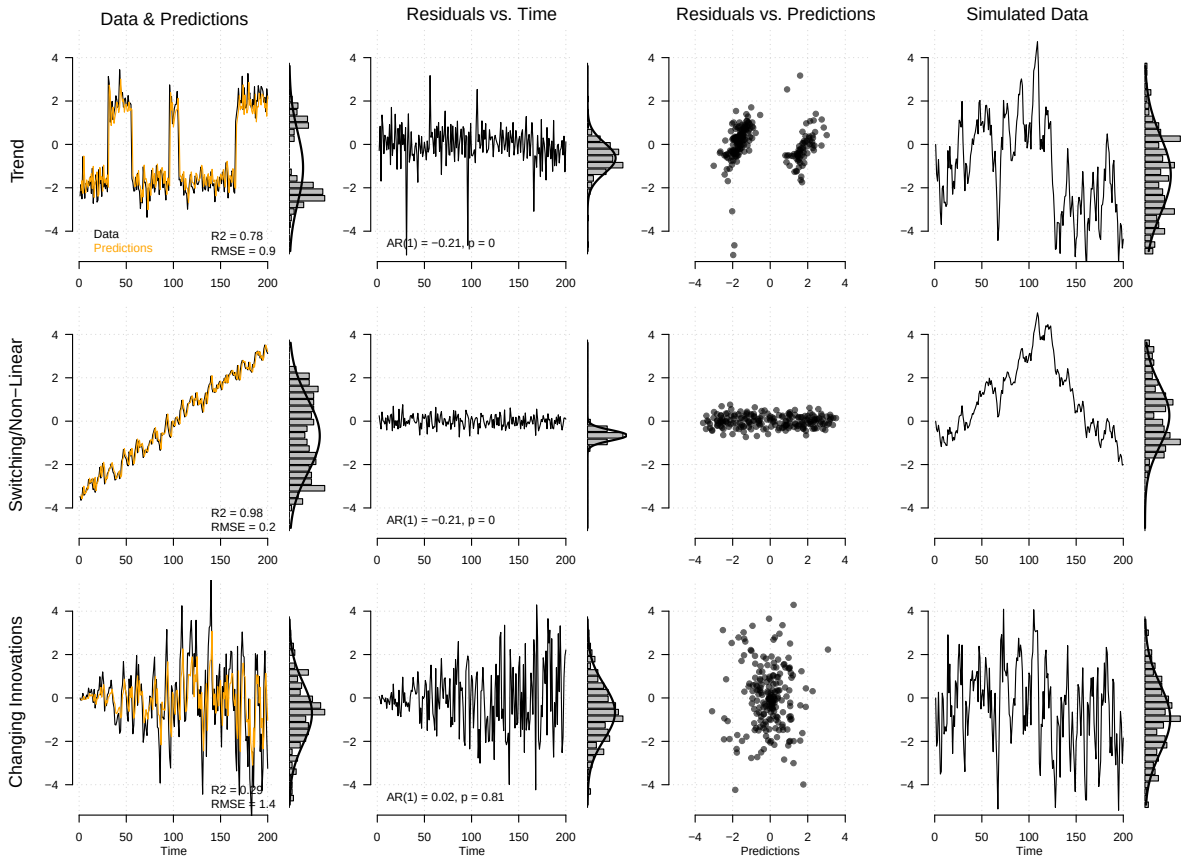


Figure 3: Three common examples of time series for which an AR-type model is misspecified: top row: a deterministic trend, which is here linear; Middle row: a non-linear/switching type of behavior; bottom row: the distribution of innovations changes across time.

*Example 4: Deterministic Trends.* The second example is a time series with a deterministic trend (second row of Figure 3). An AR(1) model does not include terms to model any trends and therefore the model is misspecified in such situations. The first plot shows the data (black line) which seems to follow a linear trend with some variation around it. We also see that the predictions (orange) align extremely well with the observed data, with a proportion of explained variance of $R^2 = 0.98$ and a rounded RMSE of 0.28. From this plot alone, one might conclude that the model fits the data extremely well. The only indication of misspecification in this plot is that the distribution of the empirical data is looking more uniform than Gaussian, which could easily be missed if the trend is not as clear as in this illustration example.

Inspecting the residuals over time in the second plot shows that the distribution of residuals seems to be approximately Gaussian throughout the entire time series. The only indication for model misspecification is that we see a negative autocorrelation in the time series. The third plot showing the residuals as a function of the predicted values shows that the residuals have a similar distribution for all predicted values and therefore does not indicate any form of misspecification. The only indication of misspecification is again that the distribution of predicted values is more uniform than Gaussian. So far, only the distributions of the data and the predictions, and the negative autocorrelation indicate misfit. From the remaining diagnostics, one could conclude that the model is correctly specified.

For this time series, the final plot showing a time series simulated from the estimated model clearly indicates model misspecification. This is because we see that the simulated time series does not at all follow the shape of the observed data. Instead, we see an AR process with very high autocorrelation. This illustrates the power of the simulation approach: sometimes a model is doing extremely well at approximating a certain time series and indicates good model fit in some diagnostic plots, even though the model is gravely misspecified. In the present case, an AR model with a very high $\phi$ coefficient (here $\hat{\phi} \approx 0.99$) is actually able to predict very well one step ahead. However, then interpreting the

high $\phi$ coefficient, for example, as high stability of the variable would of course be incorrect, as we know that the variable is in fact not stable at all.

*Example 5: Variability Changing Over Time.* The final time series shown in the bottom row of Figure 3 seems to vary around the same value but with increased variability. Any systematic change in variability is also called heteroscedasticity. Looking at the predictions (orange), we see that they are closely aligned with the observed time series early in the time series, but less so towards its end. Across the whole time series, we observe an $R^2 = 0.29$ and RMSE $= 1.46$.

While we might have been able to eyeball in the first plot that predictions are becoming worse further into the time series, we can inspect this directly in the second plot showing the residuals over time. And indeed, we see that the distribution of residuals increases in variance across time series, indicating that the AR(1) model is misspecified. The third plot showing the residuals as a function of the predicted value does not indicate any misspecification, because the present misspecification does not relate to the mean of the process, but to its variance.

The final plot with the simulated data shows a time series with high variance, matching the overall variance of the observed time series, but does not show the systematic *increase* in variance across the time series, again suggesting that the model is misspecified.

To keep things short, we here only focused on three types of misspecification, switching behavior, deterministic trend, and changes in the innovations, which we consider to be the three most common types of misspecification of the AR/VAR models in psychological time series. In Appendix D, we discuss three additional types of misspecification: seasonality, state-dependent innovations, and non-Gaussian innovations.

# 4 Model Checking for Multilevel-VAR Model in Empirical Emotion Time Series

We now apply the model checking diagnostics introduced above to a realistic empirical application. We go in two ways beyond the AR(1) model we have been working with so far: We consider the multivariate VAR(1) model, and we consider the multilevel extension of this model, which is one of the most used ones in dynamical modeling. We first describe the empirical data we use (Section 4.1) and indicate how we estimated the multilevel VAR model and the resulting parameter estimates (Section 4.2). We then perform model checking diagnostics for several persons (Section 4.3) and also compute some aggregate statistics across persons (Section 4.4). The code to reproduce all analyses can be found at https://github.com/jmbh/VARModelCheckingPaper.

## 4.1 Empirical Dataset

We use the open dataset of Grommisch et al. (2020), who investigated individual differences in the use of emotion regulation (ER) strategies in daily life using EMA and multilevel latent profile analysis to explore how the variety and combinations of ER strategies relate to well-being. They did not investigate statistical relationships across time steps with a VAR model, but their data is representative of many EMA studies that do. Their sample included 179 adults aged 18 – 69 years. The EMA design consisted of measurements over a period of 21 days during which the participants were prompted between 10 a.m. and 10 p.m. at intervals of $80 \pm 30$ minutes, resulting in approximately nine prompts per day.

During each EMA prompt, researchers assessed a number of momentary emotions and the use of emotion regulation strategies. In this tutorial, we focus on the measurements of the emotions Happy, Relaxed, Sad, and Angry. The emotion items asked how [emotion] participants felt at the moment at which the measurement was taken, and the responses were scored on a Visual Analogue Scale ranging from 0 (not at all) to 100 (very much), which was initialized at the middle of the scale at 50.

We chose this dataset because it is openly available, has a reasonably large number of participants and time points, and includes affective measurements that are often used in EMA research. From informally inspecting many time series ourselves, we were also able to judge that the present dataset is quite representative of the data from typical EMA studies.

## 4.2 Estimating Multilevel VAR Model

### 4.2.1 Preprocessing Data

We take the data as provided in the reproducibility archive by Grommisch et al. (2020). We only inserted rows of NAs where measurements were missing to properly display the missingness in time

series plots. This has no consequence for model estimation: The software uses the day number and the notification number on each day to ensure that only those time points per day are used for estimation for which the previous time point has been measured. The latter is standard practice and is necessary to obtain meaningful parameter estimates.

### 4.2.2 Estimating VAR(1) Model

We estimate the multilevel VAR(1) using the method implemented in the R-package *mlVAR* (Epskamp & Bringmann, 2017; Epskamp et al., 2018). This method estimates the multilevel VAR model by fitting a multilevel linear regression model on each response variable with all variables at the previous time point as predictors. Instead of modeling latent means, the method within-person standardizes all time series and adds the within-person means as person-level predictors. This creates known biases but is computationally efficient (for more details see Epskamp et al., 2018; Haslbeck & Epskamp, 2024).

Here we use the *mlVAR* package to estimate the multilevel VAR model because it is freely available, easy to use, and many researchers are familiar with it. However, there are many other methods to estimate multilevel VAR methods, such as the DSEM framework in M*plus* (Asparouhov et al., 2018; McNeish & Hamaker, 2020), which fits the multi-level VAR model as a joint model. Li et al. (2022) compares estimating multilevel VAR models in Mplus, Stan, and JAGs. Recently, Jongerling et al. (2024), Koslowski et al. (2024a), and Siepe, Kloft, et al. (2025) also implemented a Bayesian method to estimate multilevel VAR models using Stan. All of the model checking discussed below can equally be done with the output of any other method to estimate a multilevel VAR model, or in fact, *any* other time series model.

### 4.2.3 Inspecting Parameter Estimates

Here we inspect the parameters of the Gaussian distribution modeling the means (i.e., group average, fixed effects) of parameters and their variation across people (i.e., random effects variance) in the multilevel model. The means, or fixed effects estimates, of the lagged effects are shown on the left panel of Figure 4.

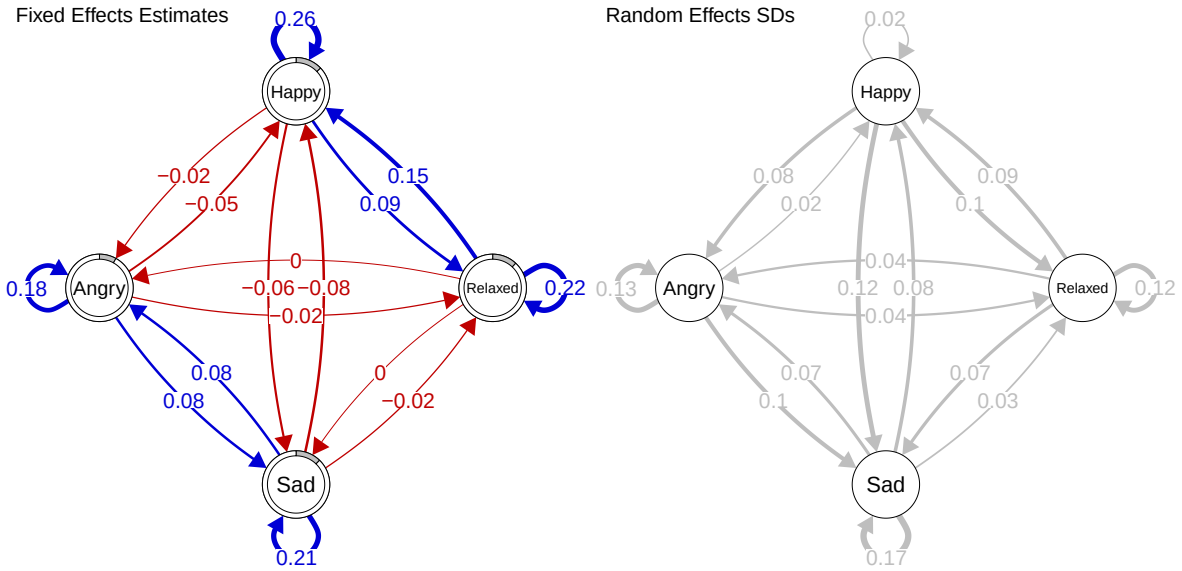


Figure 4: Left: Fixed effects estimates of lagged effects parameters. The rings around the nodes indicate within-sample the proportion of variance ($R^2$) explained for the respective variable; Right: Random effects standard deviation for each lagged parameter.

We see that there are positive autocorrelation effects between 0.18 and 0.26 and cross-lagged effects between $-0.06$ and 0.09. Cross-lagged effects are positive between emotions with the same valence and negative between emotions with different valence. These qualitative findings are typical for VAR models estimated on subjective affective time series (O. Ryan, Dablander, & Haslbeck, 2023). We also show the median proportion of variance explained ($R^2$) in each variable averaged across persons as a ring around the nodes (Haslbeck & Waldorp, 2018). We use the median because $R^2$ can become unstable if the variance of a variable is close to zero, as is often the case for negative affect items. The

median $R^2s$ are 0.13 (Happy), 0.10 (Relaxed), 0.08 (Sad) and 0.04, showing that the predictive power over a time lag of $80 \pm 30$ minutes is relatively limited. We provide a detailed analysis of prediction errors and how they differ between persons in Section 4.4 and Appendix E.

The right panel of Figure 4 shows the random effects standard deviation, so the estimated standard deviation of the distribution of a given parameter across persons in the population. The standard deviations vary between 0.02 for the autocorrelation of Happy to 0.17 for the autocorrelation of Sad, indicating that there is a lot more heterogeneity between persons in the latter (for the full random effects distributions see Appendix G).

Most studies only report the fixed effects estimates shown in the left panel of Figure 4. Few studies also report the variability across persons modeled by the random effects distribution. Focusing on fixed effects is only justified if model is fitting the data well and if the random effects variance is not too large. If the model fits poorly, standard interpretations of parameters might be a mischaracterization of the data: For example, the time series showing the switching behavior or the deterministic trend shown in Section 3 led to AR models with high autocorrelation parameters. Interpreting these as a process that varies around a stable mean with high stability between time points would be clearly incorrect.

Of course, the VAR model does not consist only of lagged effects but also of intercepts. Since *mlVAR* grand-mean centers the data before estimation the fixed effects of the intercepts are very close to zero. The random effects standard deviations rounded to two decimals happens to be 0.29 for all four variables. The full distributions are shown in Appendix G).

## 4.3 Model Checking For Individual Persons

Model checking for a multilevel VAR model involves checking how well the model fits each individual time series and whether the distribution of heterogeneity (random effects) in parameters is well modeled by a multivariate Gaussian distribution. We check the latter in Appendix G and here focus on how well the model fits each individual time series. This means that we need to inspect diagnostic plots for every variable of each of the 179 persons in the study. This means that one needs to inspect many diagnostic plots, however, we think that this is essential to understanding model fit. In rare applications in which the number of persons is prohibitively large, one could perform model checks on a random sample of persons.

The diagnostic tools introduced for the AR model in Section 3 are also the most important diagnostic tools for the VAR model. However, due to its multivariate nature there are additional diagnostics for the VAR model. For example, one can plot residuals against each predictor separately for a better diagnostic of possible unmodeled non-linear effects. In addition, many researchers model the relationships between residuals with multivariate Gaussian distribution, which implies linear relationships between residuals. The fit of this model for the residual can also be checked with residual analysis. However, since this model check is simpler than the VAR model checking and to keep this tutorial concise, we focus on checking the model fit of the VAR model using the same plots in Section 3.

Since the format of a tutorial paper does not allow us to discuss the diagnostic plots of all 179 persons, we discuss the diagnostic plots of three persons which we chose to be representative of the misspefication of most other persons in the dataset. After that, we discuss different strategies for how to deal with model misfit.

### 4.3.1 Model Checks for Person 6

We show the diagnostic plots discussed above for the four modeled variables Happy, Relaxed, Sad, and Angry for person 6 in Figure 5.

We first consider the variable Happy. The first plot shows that the data seems to roughly vary around a stable value and the distribution collapsed over time looks roughly Gaussian, indicating that the residuals might also be Gaussian distributed. We see that the model does relatively well in predicting one time step forward, with an $R^2 = 0.39$ and RMSE = 0.8 (the data here are standardized to Mean = 0, SD = 1, so predicting only with the mean would give RMSE = 1). Inspecting the residuals over time, we see that they are uncorrelated, and the distribution is Gaussian and does not change across time, as assumed by the VAR model. In the third plot, we do not see any relationship between the residuals and the predicted values. Finally, the simulated data from the estimated model looks relatively similar to the empirical time series, ignoring the fact that the simulated data has no missing observations (but we could add those to aid the visual comparison). This suggests that the VAR model is a plausible data generating mechanism and therefore correctly specified.

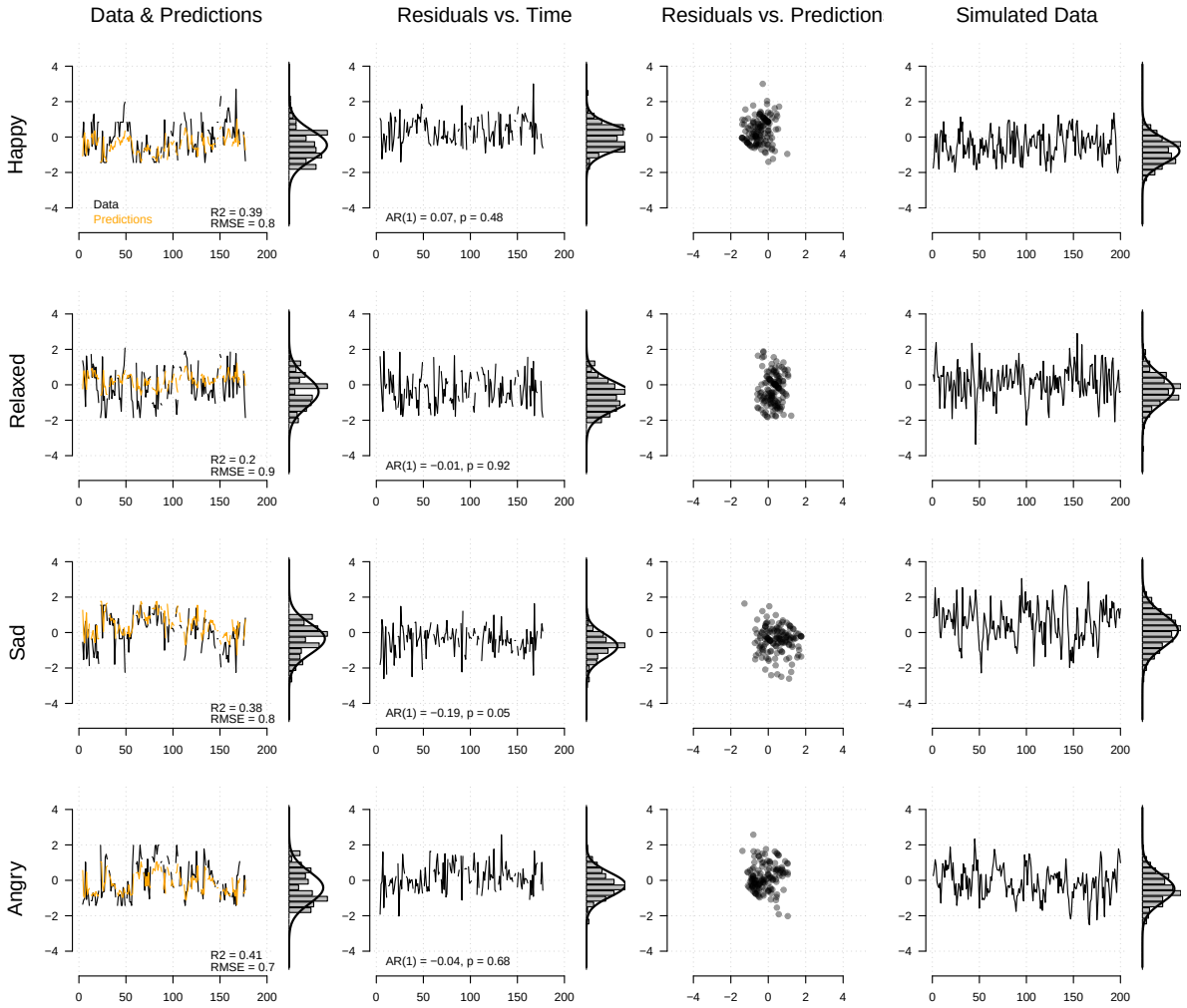


Figure 5: Model diagnostics, separately for the four modeled variables, for person 6.

We reach similar conclusions for the remaining variables Relaxed, Sad, and Angry. The residuals are roughly Gaussian distributed, they don't change over time, and they are uncorrelated, perhaps only with the exception of Sad, where we observe a negative autocorrelation of −0.19. The residual vs. predictions plots do not indicate any non-linear relationships, and the simulated data looks similar to the empirical time series. For this person, we would therefore conclude that the structure in the data is well modeled by a VAR model.

### 4.3.2 Model Checks for Person 133

Figure 6 shows the diagnostic plots for person 133. Focusing on the variable Happy, the first plot shows that there is a lot of variation up to time point 75 when the time series stabilizes at a value a bit above 0. Collapsing the observations across time we see that the distribution is skewed. Comparing observations with predictions, we find a $R^2 = 0.29$ and a RMSE = 0.92, but we see that predictions are further off the observed time series early in the time series, which already tells us that the distribution of residuals cannot be the same across time. We see this explicitly in the second plot, which also shows us that the residuals are not Gaussian distributed since there are more extreme residuals than one would expect under a Gaussian distribution. These results clearly show that the VAR(1) model is misspecified for the data of Person 133. The plot showing residuals as a function of predictions looks clearly different from the pattern we saw in the examples of correct specification in Figure 2, indicating model misfit. However, the plot does in this case not provide more information than the residual plot across time. Finally, the simulated time series does not show the clear structure of the empirical data, as we would expect when the model is gravely misspecified.

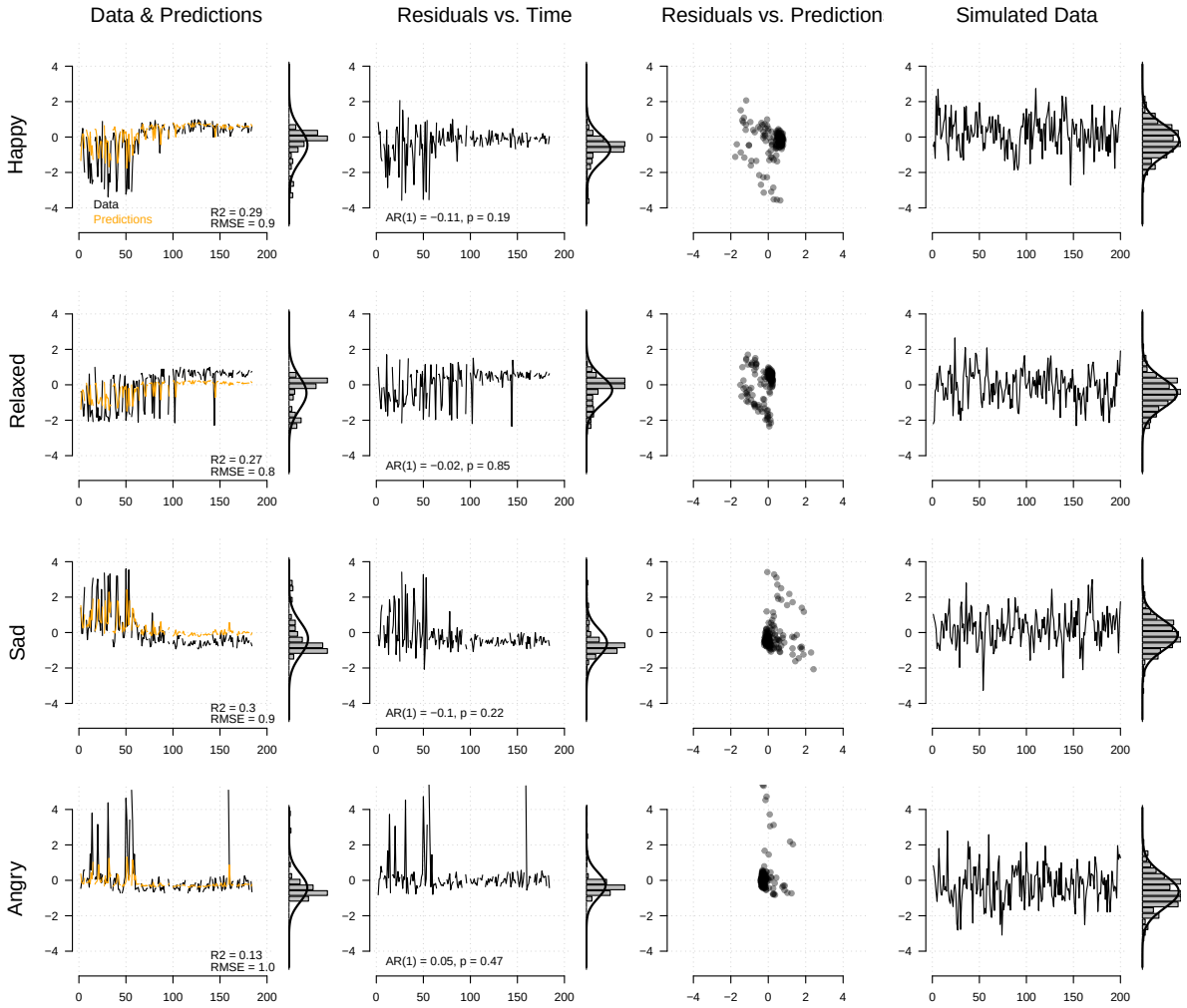


Figure 6: Model diagnostics, separately for the four modeled variables, for person 133.

Inspecting the diagnostic plots for the remaining three variables Relaxed, Sad, and Angry tells a similar story. In the first half of the time series we see periods with higher variability, whose length varies between the variables. This leads to non-Gaussian residuals and residuals that change over time. While for the variables Happy and Sad we see skewed distributions, for the variables Relaxed and Angry we see bimodal distributions with either being reasonably relaxed or not relaxed at all, or not angry or really angry. For all variables, the simulated data makes clear that the VAR(1) model is not able to capture two key aspects of the time series of the four variables, which is a kind of switching behavior between states of different shapes (depending on the variance) and that the second half of the time series has more positively valenced states.

### 4.3.3 Model Checks for Person 33

The diagnostic plots of person 33 are shown in Figure 7. We again first focus on the diagnostic plots for the variable Happy. The first plot shows that the variable seems to be switching frequently between states around $-1.5$ and 1.5. As a consequence, collapsing the observations across time leads to a bimodal distribution. The predictions are far off the observations, which is reflected by an $R^2$ close to zero, and a RMSE equal to 1. The time series of Person 33 is similar to the simulated time series with switching behavior shown in Figure 2. In the simulated switching example, the AR model predicted well, because there were few switches, which means that predicting the last time point (with $\phi$ close to 1) generally leads to very good predictions. However, in this empirical time series, the switches occur very often, which means that predicting the previous time point very often led to poor predictions. The residuals look similar to the data, only a bit less bimodal. The plot showing residuals as a function of predictions suggests some misspecification, but from this plot we do not learn more than we already know from the earlier plots. The simulated data does not reproduce the bimodal distribution of the data, indicating misspecification.

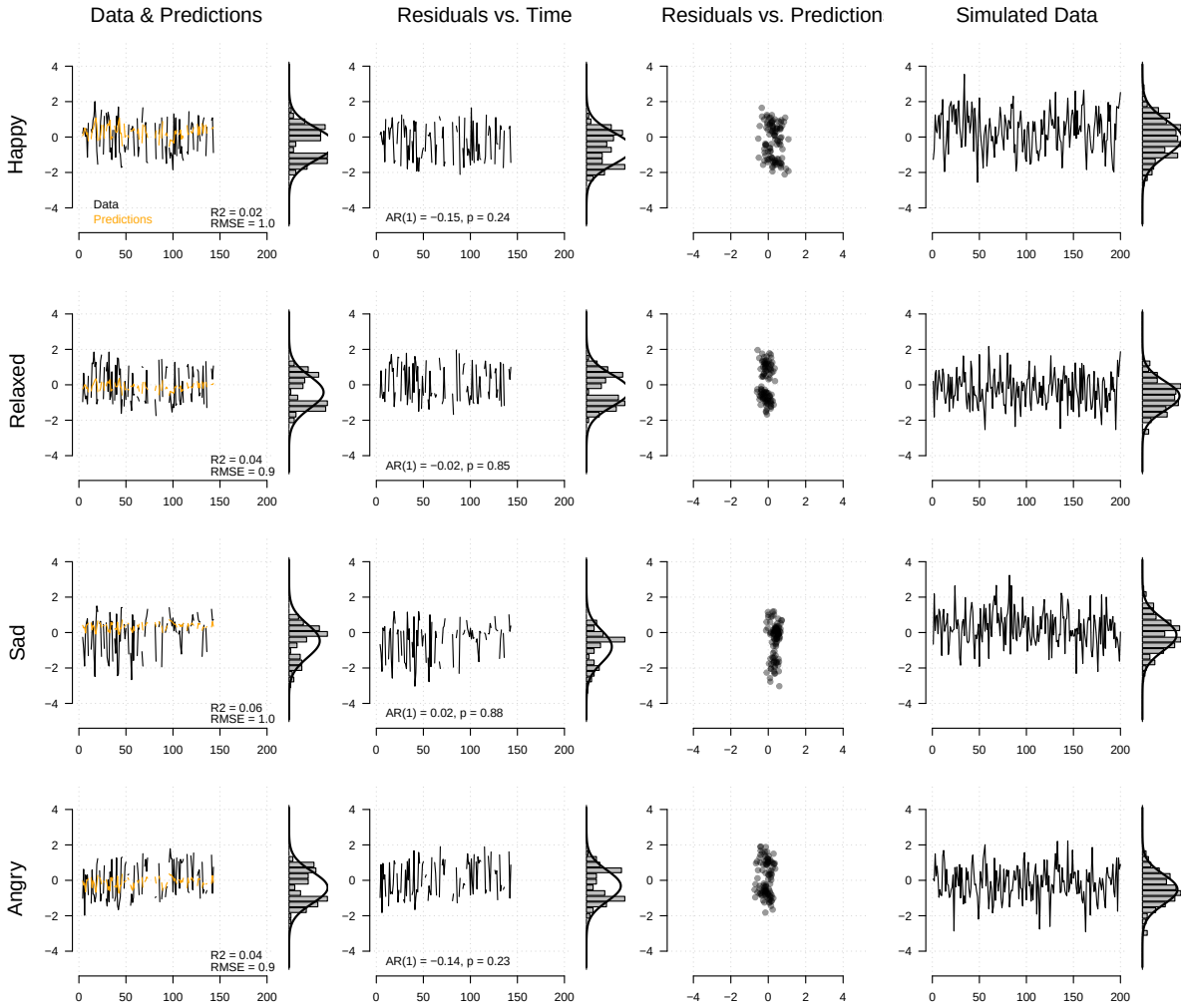


Figure 7: Model diagnostics, separately for the four modeled variables, for person 33.

Inspecting the diagnostic plots for the remaining variables Relaxed, Sad, and Angry leads us to similar conclusion. The distributions of the data are bimodal as are the residuals. The main characteristic of the time series of Person 33 is the frequent switching between states and the implied bimodality in distributions. This is a dynamic the VAR model cannot capture and the model is therefore misspecified for these data.

### 4.3.4 Overview of Model Checks for Remaining Persons

Above we provided a detailed discussion of the diagnostic plots for three selected persons. Here we summarize the impression we got from visually inspecting the diagnostic plots for the remaining 175 persons. There was a small proportion of persons for which the VAR model fit looks reasonable (e.g., persons 6, 50, and 61). However, there is also a large proportion of persons who show a switching behavior like Persons 133 and 33, indicating poor model fit (e.g., persons 32, 36, and 38). While switching is pervasive in many persons, it is not always clear whether they switch between two or more states (e.g., persons 51, 143, and 151). There is also a smaller number of persons who answer only on the lower end of the scale with occasional responses across the rest of the scale (e.g., persons 70 and 46). While it is difficult to distinguish between trends and switching more/less in a given state throughout the time series, we did not observe too many obvious trends that are not a form of switching between states (see person 55 for an exception). Figures like the ones shown above for all 178 persons can be found at https://github.com/jmbh/VARModelCheckingPaper.

## 4.4 Model Checks with Aggregate Statistics

So far, we discussed diagnostic plots which give us a detailed picture of model misfit on the person level. However, it can also be useful to assess model misfit on a more aggregate level. For example, we can estimate a linear trend for a given variable across all persons and assess whether the distribution

of slopes indicates model misfit. Such an aggregate analysis could indicate that slopes are larger than expected by chance, which might motivate an extension to model these trends in one way or another. Here we provide a couple of examples for such aggregate level model checks.

Aggregate model checks can be computed based on any quantity of interest. To continue the example from above, we might be interested in the distribution of linear trends in the data. Even if trends are non-linear, a linear trend can indicate a certain tendency of the time series to increase or decrease. We predicted for each person the time series of each variable with time and tested the null hypothesis that the slope is equal to zero with the HAC-robust test taking into account that observations are correlated (Newey & West, 1987) with $\alpha = 0.05$. Across persons and the four modeled variables the proportion of significant ($\alpha = 0.05$) linear time trends was 0.33. This is much higher than the false positive rate of 0.05 we would expect when there is no time trend, such as in a VAR model. We can also verify this by performing the same tests on the corresponding data simulated from a VAR model, which gives us a proportion of 0.06, close to the expected false positive rate. These results strongly suggest that there is at least one form of model misspecification. However, how to best address this model misfit requires careful analysis. For example, based on what we know from the diagnostic plots above, simply adding a time trend to the VAR model will not fix all issues with misspecification due to the pervasive switching behavior.

In addition, based on the observation of bimodal distributions, we might be interested in the proportion of distributions that are not unimodal (but have 2 or more modes). We determine this for each person and variable separately with Hartigan's dip test (Hartigan & Hartigan, 1985) with $\alpha = 0.05$. In the empirical data this proportion was 0.392, in the simulated data only 0.01, indicating a huge source of model misspecification in this respect. The comparison against the simulated data is especially useful when there is no test readily available for the statistic of interest. This is because it provides a formal test by repeatedly simulating datasets and using those to construct the sampling distribution of the statistic of interest under the null hypothesis that the model is correctly specified.

This section showed that we can go beyond diagnostic plots by computing quantities of interest across persons and use data simulated from the estimated model as a comparison for what we would expect under correct model misspecification. This approach is complementary to the diagnostic plots. They can be inspired by them, as in the case of multimodality, and they can go beyond them, for example by finding smaller sources of misspecification, which would be hard to spot through eyeballing diagnostic plots alone.

While we have seen in Section 3 that predictive performance is a poor proxy for model fit, it is still interesting to evaluate it and how it differs between persons. The median $R^2$s are 0.13 (Happy), 0.10 (Relaxed), 0.08 (Sad) and 0.04 (Angry), showing that for a typical person, the predictive performance of a VAR(1) model is low. In Appendix E we report the distributions of both $R^2$ and RMSE and their relationship. An (average) prediction error of VAR models is interesting but seldom reported — despite the fact that it can be computed easily and reported within network plots (see Figure 4). Note that the prediction errors we provide in Section 3 were calculated on the same data we used to fit the model. To obtain an unbiased estimate of the prediction errors in new data sampled from the same population/model, we need to compute prediction errors on (new) data we did not use for estimation (e.g., Hawinkel et al., 2024).

## 4.5 Reporting of Model Checking

How to report model checking in empirical articles? We recommend providing diagnostic plots like the ones we showed in this paper for every variable and person in an online supplementary material, which provides full transpacency and makes the model fit relatively easy to check for the reader. The results section of the paper could provide an overview of the model fit as we did in Section 4.3.4, motivating the adopted strategy to deal with potential model misfit. Additionally, one could report the distributions of key statistics that are relevant for a given dataset and compare them with the distributions expected when the model is correctly specified (see Section 4.4).

## 4.6 Conclusions from Model Checks

What are the key insights we can take away from the model checking? We saw that for some persons the model fit of the VAR(1) model was acceptable. However, we also saw that there is a large proportion of persons that show some form of a switching behavior between different "states" for which the VAR model fits poorly. We further quantified this type of behavior in Section 4.4 by looking into the proportion of bimodal distributions that are often implied by switching behavior. This confirmed that the data contains much more of this structure than one would expect by chance, given that the model

is correctly specified. We also saw that there is a considerable proportion of persons who have at least some variables on which they only score at the scale end with only occasional values throughout the rest of the scale. We could have also further quantified this, for example, by evaluating how many people / variables have a large percentage of their responses on a single value.

Overall, these results show that the VAR(1) model is misspecified for a considerable fraction of persons in the dataset. If the VAR model is the only way we explore this data, this would mean that we would miss considerable structure in the dataset, both in terms of the within-person dynamics, but also the structure of the heterogeneity between people. If we were to interpret the fixed effects estimates as a summary that well-describes the behavior of most persons, we would make a mistake. If we were to interpret the fixed effects parameters in the context of VAR-dynamics, we would also make a mistake. For example, as we know from the model checks we performed, the VAR-coefficients do not describe the dynamics of returning to a single stable state after external perturbations (i.e. innovations), which would be the case if the model was correctly specified. Instead, we have seen that this might be true for some persons, but for many others it seems to be the case that they are switching between multiple stable states.

For the present dataset, it would therefore be highly problematic to simply report the fixed effects estimates of the VAR(1) model. One could report the model misfit as context — but in the present case this would mean that we cannot interpret the VAR(1) parameters beyond them not being zero implying that we can predict one time step forward to some extent for at least some persons. Without extending the VAR model or choosing a different model (see Discussion), the fixed effects parameters can only be meaningfully be interpreted if fitted to the subset of persons for which the model fit *is* acceptable. This approach might be acceptable if only a few individuals are removed. However, removing a substantial number of individual leads to the issue that we remove persons that are systematically different from our sample, which raises the question whether the results based on the remaining persons still allow us to generalize to the initially sampled population of persons. In general, it would be preferable to avoid this issue and instead extend the (ml)VAR model to improve the model fit for (almost) all persons in the original sample. In the discussion, we review available extensions that address the types of misspecification we discussed in Section 3 and observed in Section 4.

We would expect that the model checks of VAR models would look similar in many empirical datasets. What to do when a VAR model does not fit well for many persons is of course an extremely broad question that connects back to why the VAR model was fit in the first place. However, in the discussion below we explore a couple of ideas including getting a better handle on measurement issues, different model extensions to fit additional structure in the data, and a general shift away from off-the-shelf models and towards a process of iterative model building.

# 5 Discussion

In this paper, we provided a tutorial on how to perform model checking for (multilevel) Vector Autoregressive models. We introduced the theory behind model checks, illustrated the most common forms of VAR model misspecification, and how to diagnose them with different diagnostic tools. We then went beyond simulated examples and showed how to check the fit of a VAR model in a typical empirical emotion time series dataset. Here we summarize key points and discuss complementary research areas that can contribute to better time series modeling in psychology.

## 5.1 Importance of Model Checking

We have shown that being unaware of model misfit is a problem: It can lead to missing theoretically relevant systematic effects in the data and it can lead to incorrect interpretation of the VAR parameters. However, model checking for multilevel VAR models can be daunting. This is because for VAR models there are more sources of misspecification than for standard regression models; there are a number diagnostic checks that need to be performed on each variable; and finally, all of this has to be done for a potentially large number of persons. We therefore started this tutorial with relatively simple simulated examples to walk the reader through the most common forms of VAR model misfit and how to diagnose them, before moving on to the more involved task of checking the model fit of a multilevel VAR model in empirical data.

To make it easier for researchers to perform the same analyses on their own data, we provide a well-documented code of our analysis, such that researchers can relatively easily adopt it to their own data analysis. Coding up a residual analysis involves obtaining predictions and subtracting them from the correctly preprocessed data and then making plots. Especially the former can be involved without solid knowledge of the statistical software package at hand. For that reason, we adapted the

R-package *mlVAR* such that it now returns residuals by default. And we repeated the entire analysis with DSEM/M*plus* so that readers also have reproducible code for how to perform residual analyses within this framework. The code to reproduce both analyses and all results shown in the paper can be found at https://github.com/jmbh/VARModelCheckingPaper.

While we showed that prediction errors are a poor proxy for model fit, we computed the distribution of different predictive performance measures across persons, showing that it was overall low and varied considerably between persons. We think that predictive performance is in almost all situations an interesting metric next to model fit. We therefore suggest that it should be reported, both as distributions across persons and variables, and perhaps as averages across persons within the typical network visualisation, as it is already frequently done in the context of cross-sectional network models (Haslbeck & Waldorp, 2018).

An important point we raised at the beginning of the paper was that the extent to which diagnosing different forms of model misfit is possible depends on the estimation error. If the residual is dominated by estimation error, it will be all but impossible to use it to diagnose systematic sources of model misfit. This highlights the importance of estimating VAR models with sufficient data (Mansueto et al., 2023; Zhang et al., 2025). If we estimate VAR models with insufficient data, we do not only get poor parameter estimates, but it is also very difficult to check whether the model actually fits the data.

## 5.2 Three Ways to Improve Time Series Modeling

The model checks in our empirical example showed systematic and widespread model misfit across persons. We think this is likely the case for many datasets, based on working with many empirical datasets, but also because it is plausible that the inner workings of humans are more complicated than a VAR model and that measurements taken from them consequently contain systematic effects that cannot be captured by this model. In what follows, we discuss three complementary areas of research that can contribute to better time series modeling in psychology.

### 5.2.1 Improving Measurement

One important task is to disentangle, as much as possible, effects of measurement and the latent dynamics we are primarily interested in. When estimating dynamic models such as the VAR model, the general assumption is that we are modeling statistical relationships that should give us clues about the underlying causal dynamical system we are studying. However, currently we do not know which systematic effects we commonly see in psychological time series are reflective of the dynamics unfolding between the (latent) variables of interest and which are due to measurement issues.

One prominent example is the switching behavior between different "states". This type of behavior may reflect the experience of an individual with a very volatile emotional life or/and with little differentiation for the intensity of emotions, leading to only two or a handful of discrete states a person is experiencing. However, it could also be induced by the measurement instrument, for example by initializing a slider in the middle of a VAS scale as in our study: this could nudge people to move away from zero either to the left or to the right but not staying in the middle, leading to a time series showing a kind of switching behavior and the implied multimodal distributions. However, the judgement on this and many other measurement questions is still out: despite the center-initialization of the VAS in our study, not everyone shows bimodal distributions. And conversely, also in studies without this initialization we see considerable proportions of bimodal distributions (Haslbeck et al., 2023).

This is an example of the situation in which we observe a given systematic effect (e.g., bimodality) and would like to know whether it is explained by the subjective experience of people or whether it is explained by the measurement process. However, improving measurement might also lead to the discovery of *additional* systematic effects in time series data.

Recent work is trying to better understand measurement in EMA designs using both quantitative and qualitative methodologies (e.g., Cloos et al., 2025; Haslbeck et al., 2025; Henninger et al., 2025; Schorrlepp et al., 2025). This type of research could inform design choices that ensure that everyone understands the question and uses the answer format in a similar way, which would remove a lot of systematic within- and between-person effects. This would make the modeling part easier, since it can focus on the dynamics of the latent variables instead of having to jointly model them with measurement models that differ across persons.

### 5.2.2 Embracing Model Building

Model checking tells us not only whether a model fits a given time series well, but also about the nature of a potential misfit. This allows us to adapt the model to improve model fit and gain additional insight

about the data. However, this is only possible if researchers use methodology that allows them to build models in a flexible way and if the required modeling options are available, tested, and documented. As we will see, many modeling options are already readily available for researchers, but also additional methodological work is needed.

*Using Flexible Modeling Frameworks.* Iterative model building is only possible with software that allows for a sufficiently flexible model specification. For example, the popular *mlVAR* package (Epskamp et al., 2018) is easy to use but allows one only to include or exclude (correlated or uncorrelated) random effects but otherwise does not provide any model extensions to improve model fit. Other frameworks provide more flexibility: The DSEM framework in M*plus* (Asparouhov et al., 2018; McNeish & Hamaker, 2020) supports extensions such as modeling deterministic trends, modeling not only the location (means) but also the scale (residual variances) of variables, and modeling ordinal responses (McNeish et al., 2024). Additional features of DSEM are that it allows of latent mean centering, which avoids known biases to VAR parameters, it naturally handles missing data, can be extended with measurement models, and has a method for dealing with unequal time intervals. A major downside of M*plus* is that it is not freely available. There are multiple free alternatives, but they either provide less flexibility or are more difficult to specify for researchers without technical training. A recent free alternative that is still relatively easy to specify is the R-package *mlts* (Koslowski et al., 2024b), which allows one to estimate DSEM models including random effects and time-varying covariates. A more flexible open-source option is the Bayesian multilevel package *brms* (Bürkner, 2017), which has great flexibility including the modeling of non-linear effects, almost any link function (e.g., Gaussian, Logit, Beta), location-scale modeling, and multivariate responses that allow for cross-equation residual and random effects covariances. However, constructing VAR models with *brms* requires more coding and expertise. This is true in particular for specifying uninformative priors, for which there is currently little guidance. Other open source options for estimating multilevel VAR models are *Stan* or *JAGS* (Faleh et al., 2025; Li et al., 2022), however, specifying these models requires even more expertise than *brms*.

Most of the frameworks suggested above are Bayesian. A downside of Bayesian inference is that the computational cost increases dramatically when estimating a complex model (e.g. many random effects and random effects correlations) model on more than a few variables. Computational cost for a given model and set of variables can be reduced by choosing Bayesian estimation procedures with lower cost such as variational inference (Blei et al., 2017). In some cases also Frequentist alternatives can be used, such as the R-package *lme4* (Bates et al., 2015) which has a functionality similar to *brms* and is being used within the *mlVAR* package.

A large number of model extensions are in principle already available in implementations. In what follows, we focus on the ones that are most suitable to address the three most common types of misspecification we discussed in Section 3.

*Modeling Trends.* Deterministic trends are likely the most studied type of model extension for VAR models. Time-varying VAR models have been available for some time but implementations were limited to non-hierarchical VAR models (e.g., Adolf et al., 2017; Bringmann et al., 2017; Haslbeck et al., 2021; O. Ryan et al., 2025). Linear or non-linear time trends would be straightforward to add in flexible modeling frameworks like M*plus* or *brms*. Recent simulation studies show that both cyclical (cosinor) and smooth non-linear (spline) trends are recoverable under realistic intensive-longitudinal designs and that ignoring trends biases dynamic (AR/VAR) parameters (Muthén, Asparouhov, & Keijsers, 2025a; Muthén, Asparouhov, & Shiffman, 2025; Sørensen & McCormick, 2025). Another type of deterministic trend are cyclic trends. M. Haqiqatkhah and Hamaker (2025) studied the cosinor model to model the effect of circadian rhythms and M. Haqiqatkhah and Hamaker (n.d.) extended the model to the multilevel context. While these models do not model dependencies between variables, Muthén, Asparouhov, and Keijsers (2025b) recently extended DSEM with this type of cyclic trend through a transformed linear predictor, an approach that could also be used in different software. Another way to include "cyclic" trends is to include time-varying covariates that capture repeating events such as different times of the day, or days of the week. While trends are traditionally seen as trends in means, they can also be added as moderation effects for lagged effects parameters like in non-hierarchical time-varying VAR models (e.g., Bringmann et al., 2017, 2024; Haslbeck et al., 2021). Next to deterministic trends there are also stochastic trends, however, it is currently unclear to what extent these can be modeled given the multilevel setting and the bounded scales used for almost all measurements (O. Ryan et al., 2025).

*Time-Varying Covariates and Interventions.* We mentioned that repeating events can be modeled as time-varying covariates. The effect of non-cyclic events can be modeled in a similar way. For example, we can include variables capturing events in the environment of a person or planned interventions and make means or lagged effects parameters dependent (i.e. moderated) on them (Hamaker et al., 2015; Lütkepohl, 2005).

*Modeling Non-Linear Effects.* Extending VAR models with non-linear effects between variables is interesting because the higher flexibility might lead to better model fit, but also because moderation effects might map directly on research questions, and because non-linear VAR models have more interesting theoretical properties such as multiple stable states (Strogatz, 2015). Cui et al. (2024) studied a quadratic VAR model, in which each variable is predicted by all other variables, and all other variables squared and found that a considerable number of time points are needed to support this model. It would be interesting to see to what extent this can be mitigated with a hierachical model. In addition, **failenschmid2025modeling<empty citation>** recently studied the applicability of more data-driven non-parametric approaches to model non-linear psychological processes and found that especially Generalized Additive Models and Gaussian Processes have the potential to recover underlying non-linear processes well. However, this work was on $N = 1$ time-series data, and while multilevel extensions are currently being developed, their data requirements and performance are currently unclear.

*Modeling Innovation Variances.* In the standard VAR model, the mean of each variable is modeled as a linear function of all other variables including itself at previous time points. The extensions that are modeling trends include additional time-varying predictors for the means. However, the VAR model also consists of innovations with a certain covariance matrix, which can be a function of the modeled variables or other time-varying covariates. These are forms of conditional heteroskedasticity (e.g., ARCH/GARCH, GARCH-X, or stochastic volatility) and can be implemented in location–scale frameworks (DSEM, *Stan*, *brms*) (Engle, 2001; Rast & Martin, 2021). This type of modeling is relatively rare but is theoretically highly interesting: one could, for example, study whether showing a certain behavior or having a certain experience now increases the variability in one's behaviors or experiences later in time. These models have different variants that allow different assumptions to be made about what drives within-person variability (Rast et al., 2022).

*Modeling Different Response Distributions.* The large majority of VAR models model variables as Gaussian distributions. This is despite the fact that a Gaussian is at least theoretically a bad fit for both of the commonly used response scales for EMA data (Haslbeck et al., 2025): ordinal data from Likert-scale responses are clearly not Gaussian. But also the responses from a VAS scale are not Gaussian, both because the scale is bounded (e.g. $[0, 100]$) and because the distributions are often heavily skewed (Haslbeck, Ryan, & Dablander, 2022). Alternatives are available in multilevel software allowing for location scale software, but few studies exist that study the consequences of this type of misspecification (but see M. M. Haqiqatkhah et al., 2024; McNeish & Savord, 2025). One could also model the distribution with zero-inflation, which would model the frequent phenomenon that a large proportion of responses are given at the lower scale end.

*Using Modeling Frameworks Different from VAR.* So far, we discussed relatively straightforward extensions to VAR models. However, in some cases achieving good model fit requires models that are different to (V)AR in more substantial ways. One example are Moving Average (MA) models. In these models, a variable at a given time point is not a function of its own past values, but of past random shocks (or innovations) implicitly modeling the combined impact of the environment on the modeled system. In contrast to autoregressive (AR) models, where the influence of a past shock persists indirectly through its impact on previous values, an MA model assumes that past shocks influence the current value directly but only for a limited time. Conceptually, an MA model may be appropriate if the effect of an event lasts for a fixed period and then abruptly disappears — for example, if a stressful experience influences mood for one or two days but does not produce a gradual decay over time, as would be captured by an AR-type model. AR/VAR and MA models can also be combined within the (V)ARMA framework, which allows both persistent dependencies and short-lived effects of shocks to be modeled simultaneously (Lütkepohl, 2005). Such models can be fit in DSEM, *brms* or the R-package *MTS* (Tsay et al., 2022).

In this paper, we focused only on lag-1 AR/VAR models, but many real-world psychological systems exhibit dependencies that extend further into the past. For example, effects may accumulate gradually over time — such as increasing exhaustion eventually culminating in burnout — or may unfold with

delayed or oscillatory patterns. Capturing such longer-range dependencies requires models that incorporate multiple lags. One approach is the distributed lag model (DLM), which includes predictors from several past time points, often with parameterized lag weights (e.g., geometric or spline-based decay) to provide a parsimonious description of long-run effects (Almon, 1965; Gasparrini, 2014). More general alternatives include higher-order autoregressive or vector autoregressive models (AR(p), VAR(p)), which allow the system's own past to influence its future over several time steps (Lütkepohl, 2005). State-space models and dynamic linear models (Durbin & Koopman, 2012; West & Harrison, 1997) offer even greater flexibility, enabling smooth or time-varying dynamics and latent processes that can capture both short- and long-term temporal dependencies. However, a remaining challenge to applying these models to typical EMA datasets with multiple measurements per day is how to model the night-gap between measurements (e.g., Berkhout et al., 2025). Once this challenge is addressed, we need methodological work that evaluates how well such models can be estimated with realistic datasets and tutorials that make these models accessible to researchers. Choosing different lag orders can be informed by additional model checks we did not focus on in this paper, such as the inspection of autocorrelation functions (ACFs and partial autocorrelation functions (PACFs) (Box et al., 2016).

A completely different modeling framework for time series are Hidden Markov Models (HMM; Zucchini & MacDonald, 2009), which model time series with a probabilistic mechanism switching between a number of discrete states associated with different multivariate distributions. HMMs are particularly well suited to modeling time series switching between different states, as we have observed for a subset of persons in our empirical analysis. In Appendix F we fit a HMM to the time series with the switching behavior from Section 3 and perform the same model checks to show HMM fits this time series very well. HMMs have been extended to the multilevel setting to model heterogeneity between persons and more efficient estimation (Aarts & Haslbeck, 2025a, 2025b; Mildiner Moraga & Aarts, 2024). State switching can also be combined with VAR models through Markov VAR models (Krolzig, 1997) or threshold VAR models (Hamaker et al., 2005), in which different states are associated with different VAR models.

*Model Complexity and Sample Size.* We discussed many extensions of VAR models that capture additional structure in psychological time series. This has the benefit of capturing additional structure that is theoretically interesting and to avoids that other parameter estimates are biased. However, extending the model means increasing its complexity, which has the downside that we need more data than for smaller models to obtain estimates with acceptable accuracy. The benefits of increasing the complexity of a time series model therefore need to be traded off with the larger estimation errors on all parameters, given the same sample size. Ideally, before fitting any models to empirical data, one determines through simulation whether the estimation error is acceptable with the model and research design at hand (e.g., Gelman et al., 2020). Considering practical and resource constraints on how much data can be collected might lead to the decision not to make a model more complex, even if model checks indicated that it would capture a form of misspecification or if it would be theoretically interesting. This issue points to more general limitations to the idea that all structure in time series can be captured with an exploratory statistical time series model. To make good decisions in this context we need to think in the larger context of theory development, which we discuss next.

### 5.2.3 Developing Theories

So far, we took the perspective of using a statistical model to model *all* of the structure in a time series. This is a powerful approach and is especially useful when no strong theories exist and if one opts for a more exploratory approach. At the same time, we expect various characteristics in the types of human systems we are studying, which are unlikely to be covered by statistical time series models that are feasible to be estimated from typical datasets (Haslbeck, Ryan, Robinaugh, et al., 2022). Considering the complex characteristics we theoretically often expect in psychological time series, and practical limitations to how complicated we can make statistical models, raises the question whether it is realistic to capture *all* systematic effects in time series in a single model.

Instead of using statistical time series models to capture all systematic effects in a time series, we could also use them to test and develop scientific theories. One can make predictions from a theory about a certain structure in time series data. The predictions of the theory can then be tested by evaluating whether the predicted structures do indeed show up in the empirical time series. The predicted structures could be any characteristic of a distribution (e.g., mean, variance, skewness, multimodality) and how it changes across time (e.g. not at all, gradually, sudden shifts). Another structure could be temporal relationships between variables as captured by VAR models. This was the approach taken by O. Ryan et al. (2024), who developed a computational model of emotion dynamics

which made predictions about the temporal relationships between emotion measurements; and they tested their theory empirically by comparing the VAR model predicted by their computational model with the VAR models found in the empirical literature modeling similar emotion measurements. Of course, this workflow would be most efficient if one's theory is formalized such that it can make precise predictions about such aspects (Haslbeck, Ryan, Robinaugh, et al., 2022; Robinaugh et al., 2021).

Theory construction is simplest if the theory is also a statistical model that can be directly be estimated from data. There are many examples of the latter in the area of cognitive and mathematical psychology where models are typically estimated from large numbers of experimental data (Griffiths et al., 2010; Mazur, 2006; Townsend, 2008). However, in the less controlled context of human functioning on daily life we can think of few such theories. The model proposed by O. Ryan et al. (2024) is a Markov model as a model for emotion dynamics that could be estimated as a (Hidden) Markov Model from data, but extending their simple model would likely have the cost of not being uniquely identifiable by data any more. This raises the question whether it is realistic to formulate theories about human functioning in daily life as statistical models that are feasible to estimate from realistic datasets. If not, the relationship between theory and statistical model becomes more difficult, but ideas for dealing with this situation exist (e.g., Borsboom et al., 2021; Haslbeck, Ryan, Robinaugh, et al., 2022; Robinaugh et al., 2021).

### 5.3 Conclusion

We provided a tutorial on how to perform model checks for the Vector Autoregressive model, one of the most popular models for psychological time series data. We explained why it is crucial to check model fit, introduced the basic theory of model checking, showcased how to use diagnostic plots and small simulations with the most common sources of model misfit, and applied them to check the model fit of a VAR model estimated from a typical time series of emotion measurements. We ended by discussing complementary areas that can contribute towards better time series modeling in psychology: improving measurement, embracing iterative model building, and approaching time series analysis from a theory development perspective.

We hope our tutorial makes it easier for researchers to check the fit of their (VAR) models and will help shift the norm towards making such checks a standard part of any analysis. As a consequence, we not only hope that incorrect interpretation of estimated models can be avoided, but also that researchers are inspired to venture off beaten modeling tracks and develop new models tailored to their specific research context.

*Acknowledgements.* We would like to thank Ellen Hamaker for helpful comments on an earlier version of this manuscript. We would like to thank Grommisch et al. (2020) for making their data openly available so we could reuse it in this paper. JMBH was supported by the Netherlands Organisation for Scientific Research (NWO) under VENI grant number 221G.110. JMBH and LJW have been supported by the gravitation 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).

*Materials.* The data and code to reproduce everything in this paper are available at https://github.com/jmbh/VARModelCheckingPaper.

## References

Aalbers, G., McNally, R. J., Heeren, A., de Wit, S., & Fried, E. I. (2019). Social media and depression symptoms: A network perspective. *J. Exp. Psychol. Gen.*, *148*(8), 1454–1462.

Aarts, E., & Haslbeck, J. M. B. (2025a). Modeling psychological time series with multilevel hidden markov models: A numerical evaluation.

Aarts, E., & Haslbeck, J. M. B. (2025b). Modeling psychological time series with multilevel hidden markov models: A tutorial.

Adolf, J. K., Voelkle, M. C., Brose, A., & Schmiedek, F. (2017). Capturing context-related change in emotional dynamics via fixed moderated time series analysis. *Multivariate behavioral research*, *52*(4), 499–531.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, 178–196.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural equation modeling: a multidisciplinary journal*, *25*(3), 359–388.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berkhof, J., Van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*(3), 337–354.

Berkhout, S. W., Schuurman, N. K., & Hamaker, E. L. (2025). Let sleeping dogs lie? how to deal with the night gap problem in experience sampling method data [Online ahead of print (May 22 2025)]. *Psychological Methods*. https://doi.org/10.1037/met0000762

Blanchard, M. A., Contreras, A., Kalkan, R. B., & Heeren, A. (2023). Auditing the research practices and statistical analyses of the group-level temporal network approach to psychological constructs: A systematic scoping review. *Behavior research methods*, *55*(2), 767–787.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Borsboom, D., Van Der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766.

Bos, F. M., Snippe, E., de Vos, S., Hartmann, J. A., Simons, C. J. P., van der Krieke, L., de Jonge, P., & Wichers, M. (2017). Can we jump from cross-sectional to dynamic interpretations of networks? implications for the network perspective in psychiatry. *Psychother. Psychosom.*, *86*(3), 175–177.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons. https://doi.org/10.1002/9781118619193

Bringmann, L. F., Ariens, S., Ernst, A. F., Snippe, E., & Ceulemans, E. (2024). Changing networks: Moderated idiographic psychological networks. *advances.in/psychology*, *2*, e658296. https://doi.org/10.56296/aip00014

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological methods*, *22*(3), 409.

Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the beck depression Inventory-II. *Psychol. Med.*, *45*(4), 747–757.

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment*, *23*(4), 425–435.

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.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Cloos, L., Siepe, B. S., Piccirillo, M. L., Fried, E., Wang, S. B., Helmich, M. A., Johnson, S. U., Hoffart, A., & Ebrahimi, O. V. (2025). Accuracy and consistency of visual analog scales in ecological momentary assessment and digital studies. *Collabra: Psychology*, *11*(1), 142735.

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.

Contreras, A., Valiente, C., Heeren, A., & Bentall, R. (2020). A temporal network approach to paranoia: A pilot study. *Front. Psychol.*, *11*(544565), 544565.

Cui, J., Lichtwarck-Aschoff, A., & Hasselman, F. (2024). Examining the feasibility of nonlinear vector autoregressions for psychological intensive longitudinal data.

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.

de Vos, S., Wardenaar, K. J., Bos, E. H., Wit, E. C., Bouwmans, M. E. J., & de Jonge, P. (2017). An investigation of emotion dynamics in major depressive disorder patients and healthy persons using sparse longitudinal networks. *PLoS One*, *12*(6), e0178586.

De Vos, S., Wardenaar, K. J., Bos, E. H., Wit, E. C., Bouwmans, M. E., & De Jonge, P. (2017). An investigation of emotion dynamics in major depressive disorder patients and healthy persons using sparse longitudinal networks. *PLoS One*, *12*(6), e0178586.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531

Dorais, S. (2024). Time series analysis in preventive intervention research: A step-by-step guide. *Journal of Counseling & Development*, *102*(2), 239–250.

Draper, N. (1998). *Applied regression analysis*. McGraw-Hill. Inc.

Drukker, M., Peters, J. C. H., Vork, L., Mujagic, Z., Rutten, B. P. F., van Os, J., Masclee, A. A. M., Kruimel, J. W., & Leue, C. (2020). Network approach of mood and functional gastrointestinal symptom dynamics in relation to childhood trauma in patients with irritable bowel syndrome and comorbid panic disorder. *J. Psychosom. Res.*, *139*(110261), 110261.

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199641178.001.0001

Ellison, W. D., Levy, K. N., Newman, M. G., Pincus, A. L., Wilson, S. J., & Molenaar, P. (2020). Dynamics among borderline personality and anxiety features in psychotherapy outpatients: An exploration of nomothetic and idiographic patterns. *Personality Disorders: Theory, Research, and Treatment*, *11*(2), 131. https://doi.org/10.1037/per0000363

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.

Engle, R. (2001). Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, *15*(4), 157–168.

Epskamp, S., & Bringmann, L. F. (2017). *Mlvar: Multi-level vector auto-regression* [R package version 0.4.4]. https://CRAN.R-project.org/package=mlVAR

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.

Faelens, L., Hoorelbeke, K., Soenens, B., Van Gaeveren, K., De Marez, L., De Raedt, R., & Koster, E. H. (2021). Social media use and well-being: A prospective experience-sampling study. *Computers in Human Behavior*, *114*, 106510.

Faleh, R., Morelli, S., Andriamiarana, V., Roman, Z. J., Flückiger, C., & Brandt, H. (2025). Dynamic latent class structural equation modeling: A hands-on tutorial for modeling intensive longitudinal data. *arXiv preprint arXiv:2508.12983*.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage publications.

Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., et al. (2024). So you want to do esm? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Science*, *7*(3), 25152459241267912.

Gasparrini, A. (2014). Modeling exposure–lag–response associations with distributed lag non-linear models. *Statistics in medicine*, *33*(5), 881–899.

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and other stories*. Cambridge University Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Geraets, C. N. W., Snippe, E., van Beilen, M., Pot-Kolder, R. M. C. A., Wichers, M., van der Gaag, M., & Veling, W. (2020). Virtual reality based cognitive behavioral therapy for paranoia: Effects on mental states and the dynamics among them. *Schizophr. Res.*, *222*, 227–234.

Greene, T., Gelkopf, M., Epskamp, S., & Fried, E. (2018). Dynamic networks of PTSD symptoms during conflict. *Psychol. Med.*, *48*(14), 2409–2417.

Greene, T., Gelkopf, M., Fried, E. I., Robinaugh, D. J., & Lapid Pickman, L. (2020). Dynamic network analysis of negative emotions and DSM-5 posttraumatic stress disorder symptom clusters during conflict. *Journal of traumatic stress*, *33*(1), 72–83.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Groen, R. N., Ryan, O., Wigman, J. T. W., Riese, H., Penninx, B. W. J. H., Giltay, E. J., Wichers, M., & Hartman, C. A. (2020). Comorbidity between depression and anxiety: Assessing the role of bridge mental states in dynamic psychological networks. *BMC Med.*, *18*(1), 308.

Groen, R. N., Snippe, E., Bringmann, L. F., Simons, C. J. P., Hartmann, J. A., Bos, E. H., & Wichers, M. (2019). Capturing the risk of persisting depressive symptoms: A dynamic network investigation of patients' daily symptom experiences. *Psychiatry Res.*, *271*, 640–648.

Grommisch, G., Koval, P., Hinton, J. D., Gleeson, J., Hollenstein, T., Kuppens, P., & Lischetzke, T. (2020). Modeling individual differences in emotion regulation repertoire in daily life with multilevel latent profile analysis. *Emotion*, *20*(8), 1462.

Hamaker, E. L. (2025). Analysis of intensive longitudinal data: Putting psychological processes in perspective. *Annual Review of Clinical Psychology*, *21*.

Hamaker, E. L., Ceulemans, E., Grasman, R. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate behavioral research*, *40*(2), 207–233.

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.

Haqiqatkhah, M., & Hamaker, E. L. (n.d.). Peaks and pitfalls: Unwrapping four overlooked problems in multilevel modeling of cyclic trends and their circular solutions.

Haqiqatkhah, M., & Hamaker, E. L. (2025). Hunting high and low: On the estimation and appropriateness of cosine models for experience sampling.

Haqiqatkhah, M. M., & Hamaker, E. L. (2025). Daily dynamics and weekly rhythms: A tutorial on seasonal autoregressive–moving average models combined with day-of-the-week effects. *Psychological Methods*.

Haqiqatkhah, M. M., Ryan, O., & Hamaker, E. L. (2024). Skewness and staging: Does the floor effect induce bias in multilevel ar(1) models? *Multivariate Behavioral Research*, *59*(2), 289–319. https://doi.org/10.1080/00273171.2023.2254769

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84. https://doi.org/10.1214/aos/1176346577

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. (2024). Observed correlations between person-means depend on within-person correlations. *Advances. in/psychology*, *2*, e853425.

Haslbeck, J. M. B., Martínez, A. J., Roefs, A. J., Fried, E. I., Lemmens, L. H., Groot, E., & Edelsbrunner, P. A. (2025). Comparing likert and visual analogue scales in ecological momentary assessment. *Behavior Research Methods*, *57*(8), 1–41.

Haslbeck, J. M. B., & Ryan, O. (2022). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, *57*(5), 735–766.

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022). Multimodality and skewness in emotion time series. https://psyarxiv.com/qudr6

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion*, *23*(8), 2117.

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, *27*(6), 930.

Haslbeck, J. M. B., & Waldorp, L. J. (2018). How well do network models predict observations? on the importance of predictability in network models. *Behavior research methods*, *50*(2), 853–861.

Hasmi, L., Drukker, M., Guloksuz, S., Menne-Lothmann, C., Decoster, J., van Winkel, R., Collip, D., Delespaul, P., De Hert, M., Derom, C., et al. (2017). Network approach to understanding emotion dynamics in relation to childhood trauma and genetic liability to psychopathology: Replication of a prospective experience sampling analysis. *Frontiers in psychology*, *8*, 1908.

Hawinkel, S., Waegeman, W., & Maere, S. (2024). Out-of-sample r 2: Estimation and inference. *The American Statistician*, *78*(1), 15–25.

Henninger, M., Vanhasbroeck, N., & Tuerlinckx, F. (2025). Affect dynamics or response bias? the relationship between extreme response style and affect dynamics in a controlled experiment. *Psychological Assessment*.

Hoffart, A., & Johnson, S. U. (2020). Within-person networks of clinical features of social anxiety disorder during cognitive and interpersonal therapy. *J. Anxiety Disord.*, *76*(102312), 102312.

Hoorelbeke, K., Van den Bergh, N., Wichers, M., & Koster, E. H. (2019). Between vulnerability and resilience: A network analysis of fluctuations in cognitive risk and protective factors following remission from depression. *Behaviour Research and Therapy*, *116*, 1–9.

Huckins, J. F., DaSilva, A. W., Hedlund, E. L., Murphy, E. I., Rogers, C., Wang, W., Obuchi, M., Holtzheimer, P. E., Wagner, D. D., & Campbell, A. T. (2020). Causal factors of anxiety

and depression in college students: Longitudinal ecological momentary assessment and causal analysis using peter and clark momentary conditional independence. *JMIR Ment. Health*, *7*(6), e16684.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd). OTexts. https://otexts.com/fpp3/

Johnson, S. U., & Hoffart, A. (2018). Metacognitive therapy versus cognitive behavioral therapy:a network approach. *Front. Psychol.*, *9*, 2382.

Jongeneel, A., Aalbers, G., Bell, I., Fried, E. I., Delespaul, P., Riper, H., Van Der Gaag, M., & Van Den Berg, D. (2020). A time-series network approach to auditory verbal hallucinations: Examining dynamic interactions using experience sampling methodology. *Schizophrenia research*, *215*, 148–156.

Jongerling, J., Liu, S., & Williams, D. R. (2024). Bayesian multilevel VAR models with random covariance matrices: Estimation requirements. osf.io/preprints/psyarxiv/5fk8v_v1

Kaiser, T., & Laireiter, A.-R. (2019). Daily dynamic assessment and modelling of intersession processes in ambulatory psychotherapy: A proof of concept study. *Psychother. Res.*, *29*(8), 1062–1073.

Klippel, A., Viechtbauer, W., Reininghaus, U., Wigman, J., van Borkulo, C., MERGE, Myin-Germeys, I., & Wichers, M. (2018). The cascade of stress: A network approach to explore differential dynamics in populations varying in risk for psychosis. *Schizophrenia Bulletin*, *44*(2), 328–337.

Koslowski, K., Münch, F., Koch, T., & Holtmann, J. (2024a). Mlts: Multilevel latent time series models with R and Stan. https://github.com/munchfab/mlts

Koslowski, K., Münch, F., Koch, T., & Holtmann, J. (2024b). Mlts: Multilevel latent time series models with R and Stan. https://github.com/munchfab/mlts

Kreiter, D., Drukker, M., Mujagic, Z., Vork, L., Rutten, B. P. F., van Os, J., Masclee, A. A. M., Kruimel, J. W., & Leue, C. (2021). Symptom-network dynamics in irritable bowel syndrome with comorbid panic disorder using electronic momentary assessment: A randomized controlled trial of escitalopram vs. placebo. *J. Psychosom. Res.*, *141*(110351), 110351.

Krolzig, H.-M. (1997). The markov-switching vector autoregressive model. In *Markov-switching vector autoregressions: Modelling, statistical inference, and application to business cycle analysis* (pp. 6–28). Springer.

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.

Kuranova, A., Wigman, J. T., Menne-Lothmann, C., Decoster, J., van Winkel, R., Delespaul, P., Drukker, M., de Hert, M., Derom, C., Thiery, E., et al. (2021). Network dynamics of momentary affect states and future course of psychopathology in adolescents. *PloS One*, *16*(3), e0247458.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, *54*(1-3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lazarus, G., Sened, H., & Rafaeli, E. (2020). Subjectifying the personality state: Theoretical underpinnings and an empirical example. *Eur. J. Pers.*, *34*(6), 1017–1036.

Levinson, C. A., Vanzhula, I., & Brosof, L. C. (2018). Longitudinal and personalized networks of eating disorder cognitions and behaviors: Targets for precision intervention a proof of concept study. *Int. J. Eat. Disord.*, *51*(11), 1233–1243.

Levinson, C. A., Vanzhula, I. A., Smith, T. W., & Stice, E. (2020). Group and longitudinal intra-individual networks of eating disorder symptoms in adolescents and young adults at-risk for an eating disorder. *Behav. Res. Ther.*, *135*(103731), 103731.

Li, Y., Wood, J., Ji, L., Chow, S.-M., & Oravecz, Z. (2022). Fitting multilevel vector autoregressive models in stan, jags, and mplus. *Structural equation modeling: a multidisciplinary journal*, *29*(3), 452–475.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, *8*(1), 1–9.

Mansueto, A. C., Wiers, R. W., van Weert, J., Schouten, B. C., & Epskamp, S. (2023). Investigating the feasibility of idiographic network models. *Psychological methods*, *28*(5), 1052.

Martín-Brufau, R., Suso-Ribera, C., & Corbalán, J. (2020). Emotion network analysis during COVID-19 quarantine - a longitudinal study. *Front. Psychol.*, *11*, 559572.

Mazur, J. E. (2006). Mathematical models and the experimental analysis of behavior. *Journal of the experimental analysis of behavior*, *85*(2), 275–291.

McCuish, E., Bouchard, M., & Beauregard, E. (2021). A network-based examination of the longitudinal association between psychopathy and offending versatility. *J. Quant. Criminol.*, *37*(3), 693–714.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

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.

McNeish, D., & Savord, A. (2025). Exploring how many categories are needed to model ordinal intensive longitudinal data as continuous with dynamic structural equation models [Online ahead of print]. *Psychological Methods*. https://doi.org/10.1037/met0000784

McNeish, D., Somers, J. A., & Savord, A. (2024). Dynamic structural equation models with binary and ordinal outcomes in m plus. *Behavior research methods*, *56*(3), 1506–1532.

Meng, J., Wang, X., Wei, D., & Qiu, J. (2020). State loneliness is associated with emotional hypervigilance in daily life: A network analysis. *Personality and Individual Differences*, *165*, 110154.

Mildiner Moraga, S., & Aarts, E. (2024). Go multivariate: Recommendations on bayesian multilevel hidden markov models with categorical data. *Multivariate Behavioral Research*, *59*(1), 17–45.

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

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Muthén, B., Asparouhov, T., & Keijsers, L. (2025a). Dynamic structural equation modeling with cycles. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(2), 264–286. https://doi.org/10.1080/10705511.2024.2406510

Muthén, B., Asparouhov, T., & Keijsers, L. (2025b). Dynamic structural equation modeling with cycles. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(2), 264–286.

Muthén, B., Asparouhov, T., & Shiffman, S. (2025). Dynamic structural equation modeling with floor effects [Advance online publication]. *Psychological Methods*. https://doi.org/10.1037/met0000720

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708. https://doi.org/10.2307/1913610

Oreel, T. H., Borsboom, D., Epskamp, S., Hartog, I. D., Netjes, J. E., Nieuwkerk, P. T., Henriques, J. P., Scherer-Rath, M., van Laarhoven, H. W., & Sprangers, M. A. (2019). The dynamics in health-related quality of life of patients with stable coronary artery disease were revealed: A network analysis. *Journal of Clinical Epidemiology*, *107*, 116–123.

Pannicke, B., Kaiser, T., Reichenberger, J., & Blechert, J. (2021). Networks of stress, affect and eating behaviour: Anticipated stress coping predicts goal-congruent eating in young adults. *International Journal of Behavioral Nutrition and Physical Activity*, *18*(1), 1–14.

Pavani, J.-.-B., Le Vigouroux, S., Kop, J.-.-L., Congard, A., Dauvier, B., & Denissen, J. (2017). A Network Approach to Affect Regulation Dynamics and Personality Trait–Induced Variations: Extraversion and Neuroticism Moderate Reciprocal Influences between Affect and Affect Regulation Strategies. *European Journal of Personality*, *31*(4), 329–346.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., et al. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, *3*(2), 292–300.

Peeters, L., Van Den Noortgate, W., Blanchard, M. A., Suenaert, M., Eisele, G. V., Kirtley, O. J., Artner, R., & Lafit, G. (n.d.). Mapping methodological variation in experience sampling research from design to data analysis: A systematic review.

Rast, P., & Martin, S. R. (2021). Bmgarch: An R-package for Bayesian Multivariate GARCH models. *Journal of Open Source Software*, *6*(64), 3452. https://doi.org/10.21105/joss.03452

Rast, P., Martin, S. R., Liu, S., & Williams, D. R. (2022). A new frontier for studying within-person variability: Bayesian multivariate generalized autoregressive conditional heteroskedasticity models. *Psychological Methods*, *27*(5), 856–873. https://doi.org/10.1037/met0000357

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743.

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

Ryan, O., Dablander, F., & Haslbeck, J. M. B. (2024). Toward a generative model for emotion dynamics. *Psychological review*.

Ryan, O., Haslbeck, J. M. B., & Waldorp, L. (2023). Non-stationarity in time-series analysis: Modeling stochastic and deterministic trends.

Ryan, O., Haslbeck, J. M. B., & Waldorp, L. J. (2025). Non-stationarity in time-series analysis: Modeling stochastic and deterministic trends. *Multivariate Behavioral Research*, *60*(3), 556–588.

Ryan, T. P. (2008). *Modern regression methods* (Vol. 655). John Wiley & Sons.

Schorrlepp, L., Stadel, M., Bringmann, L. F., Hesselink, M., & Maciejewski, D. (2025). Utilizing qualitative methods to detect validity issues in clinical experience sampling methodology (esm).

Siepe, B. S., Kloft, M., Zhang, Y., Petersen, F., Bringmann, L. F., & Heck, D. W. (2025). Using features of dynamic networks to guide treatment selection and outcome prediction: The central role of uncertainty.

Siepe, B. S., Rieble, C. L., Tutunji, R., Rimpler, A., März, J., Proppert, R. K., & Fried, E. I. (2025). Understanding ecological-momentary-assessment data: A tutorial on exploring item performance in ecological-momentary-assessment data. *Advances in Methods and Practices in Psychological Science*, *8*(1), 25152459241286877.

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), 1–10.

Sørensen, Ø., & McCormick, E. M. (2025). Modeling cycles, trends and time-varying effects in dynamic structural equation models with regression splines. *Multivariate Behavioral Research*, *60*(5), 1013–1028. https://doi.org/10.1080/00273171.2025.2507297

Strogatz, S. (2015). Nonlinear dynamics and chaos 2nd edn (boulder, co.

Thonon, B., Contreras, A., & Larøi, F. (2022). Motivation in schizophrenia: Preliminary findings of a theory-driven approach using time-series network analysis. *Curr. Psychol.*, *41*(11), 7731–7741.

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of mathematical psychology*, *52*(5), 269–280.

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

Tsay, R. S., Wood, D., & Lachmann, J. (2022). *Mts: All-purpose toolkit for analyzing multivariate time series (mts) and estimating multivariate volatility models* [R package]. Version 1.2.1. https://cran.r-project.org/package=MTS

van Roekel, E., Heininga, V. E., Vrijen, C., Snippe, E., & Oldehinkel, A. J. (2019). Reciprocal associations between positive emotions and motivation in daily life: Network analyses in anhedonic individuals and healthy controls. *Emotion*, *19*(2), 292–300.

Veenman, M., Janssen, L. H., van Houtum, L. A., Wever, M. C., Verkuil, B., Epskamp, S., Fried, E. I., & Elzinga, B. M. (2024). A network study of family affect systems in daily life. *Multivariate Behavioral Research*, *59*(2), 371–405.

Visser, I., & Speekenbrink, M. (2010). Depmixs4: An r package for hidden markov models. *Journal of statistical Software*, *36*, 1–21.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer. https://doi.org/10.1007/978-0-387-21736-9

West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-9365-9

Wigman, J. T. W., van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., MERGE, Viechtbauer, W., Myin-Germeys, I., & Wichers, M. (2015). Exploring the underlying structure of mental disorders: Cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. *Psychol. Med.*, *45*(11), 2375–2387.

Zhang, Y., Revol, J., Lafit, G., Ernst, A., Razum, J., Ceulemans, E., & Bringmann, L. F. (2025). Meeting the bare minimum: Quality assessment of idiographic temporal networks using power analysis and predictive accuracy analysis.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden markov models for time series: An introduction using r*. Chapman; Hall/CRC.

# A Small Review of VAR Model Checking in the Literature

We conducted a brief review of model checking techniques in published papers using group-level temporal networks, predominantly multilevel vector autoregression, based on the literature review by (Blanchard et al., 2023). We used the list of 43 articles they found and already coded for stationarity checks. We additionally reviewed them for any form of residual model check. We summarize our results in Table 1. Overall, four studies (9.3%) performed some residual model check, none explicitly checked for heteroskedasticity, and 29 (67.4%) checked for stationarity in some way.

| Author | Residuals | Hetero-skedasticity | Stationarity | Stationarity Category |
|---|---|---|---|---|
| Aalbers et al. (2019) | ✗ | ✗ | ✓ | stochastic |
| Bos et al. (2017) | ✗ | ✗ | ✓ | deterministic |
| Bringmann et al. (2016) | ✗ | ✗ | ✗ | |
| Bringmann et al. (2015) | ✗ | ✗ | ✓ | deterministic |
| Bringmann et al. (2013) | ✗ | ✗ | ✓ | stochastic/deterministic |
| Curtiss et al. (2019) | ✗ | ✗ | ✗ | |
| de Vos et al. (2017) | ✓ | ✗ | ✓ | deterministic |
| Ellison et al. (2020) | ✗ | ✗ | ✗ | |
| Greene et al. (2020) | ✗ | ✗ | ✗ | |
| Greene et al. (2018) | ✗ | ✗ | ✗ | |
| Groen et al. (2019) | ✗ | ✗ | ✓ | deterministic |
| Hasmi et al. (2017) | ✗ | ✗ | ✓ | deterministic |
| Hoorelbeke et al. (2019) | ✗ | ✗ | ✗ | |
| Johnson and Hoffart (2018) | ✗ | ✗ | ✓ | stochastic/deterministic |
| Jongeneel et al. (2020) | ✓ | ✗ | ✓ | stochastic/deterministic |
| Kaiser and Laireiter (2019) | ✗ | ✗ | ✓ | stochastic/deterministic |
| Klippel et al. (2018) | ✗ | ✗ | ✓ | deterministic |
| Levinson et al. (2018) | ✗ | ✗ | ✗ | |
| Lutz et al. (2018) | ✗ | ✗ | ✓ | stochastic |
| Oreel et al. (2019) | ✗ | ✗ | ✓ | deterministic |
| Pavani et al. (2017) | ✓ | ✗ | ✓ | deterministic |
| Pe et al. (2015) | ✗ | ✗ | ✗ | |
| **rath2019modeling<empty citation>** | ✗ | ✗ | ✓ | stochastic |
| Snippe et al. (2017) | ✗ | ✗ | ✓ | deterministic |
| van Roekel et al. (2019) | ✗ | ✗ | ✓ | deterministic |
| Wigman et al. (2015) | ✗ | ✗ | ✗ | |
| Contreras et al. (2020) | ✗ | ✗ | ✓ | stochastic |
| Drukker et al. (2020) | ✗ | ✗ | ✓ | deterministic |
| Elovainio et al. (2020) | ✗ | ✗ | ✗ | |
| Faelens et al. (2021) | ✗ | ✗ | ✓ | stochastic |
| Geraets et al. (2020) | ✗ | ✗ | ✓ | deterministic |
| Groen et al. (2020) | ✗ | ✗ | ✓ | stochastic/deterministic/visual |
| Hoffart and Johnson (2020) | ✗ | ✗ | ✓ | stochastic/deterministic |
| Huckins et al. (2020) | ✗ | ✗ | ✗ | |
| Kreiter et al. (2021) | ✗ | ✗ | ✓ | deterministic |
| Kuranova et al. (2021) | ✓ | ✗ | ✓ | deterministic |
| Lazarus et al. (2020) | ✗ | ✗ | ✓ | deterministic |
| Levinson et al. (2020) | ✗ | ✗ | ✓ | visual |
| Martín-Brufau et al. (2020) | ✗ | ✗ | ✗ | |
| Meng et al. (2020) | ✗ | ✗ | ✗ | |
| Pannicke et al. (2021) | ✗ | ✗ | ✗ | |
| Thonon et al. (2022) | ✗ | ✗ | ✓ | stochastic |
| McCuish et al. (2021) | ✗ | ✗ | ✓ | stochastic/visual |

Table 1: Results of reviewing the model checking practices in 43 published papers reporting VAR models.

The four studies that investigated the residuals did so because they were interested in the influence of skewness or in overall model fit. None of them explicitly investigated potential heteroskedasticity. De Vos et al. (2017) mentioned non-normal residuals as reasoning for a quantile transformation, while Pavani et al. (2017) decided not to transform a skewed variable because model residuals were normally distributed. Kuranova et al. (2021) visually checked the residual distribution of all multilevel models

used in their study for normality, while Jongeneel et al. (2020) used a statistical test to do so.

We separated stationarity checks into the inspection or correction of stochastic trends via Kwiatkowski–Phillips–Schmi tests (Kwiatkowski et al., 1992), Dickey-Fuller tests (Dickey & Fuller, 1979), or differencing, the inspection or correction of stochastic trends, typically via estimating a linear effect of time, and visual inspections of time series for trends. Of the 29 studies that checked for stationarity, 21 inspected or removed deterministic trends, which was the most common form of dealing with potential trends.

# B Data Generating Models for Simulated Examples

We generated eight different data generating models (DGMs) based on autoregressive (AR) processes. All processes are defined for $t = 1, \ldots, N$ with $N = 200$. Unless otherwise specified, innovations $\varepsilon_t$ are i.i.d. Gaussian with mean zero and variance one, i.e. $\varepsilon_t \sim \mathcal{N}(0, 1)$.

### Model 1: Correctly specified, AR $> 0$

$$x_t = \alpha + \phi x_{t-1} + \varepsilon_t, \quad \alpha = 0,\ \phi = 0.6,\ \varepsilon_t \sim \mathcal{N}(0, 1).$$

### Model 2: Correctly specified, AR $= 0$

$$x_t = \alpha + \phi x_{t-1} + \varepsilon_t = \varepsilon_t, \quad \alpha = 0,\ \phi = 0,\ \varepsilon_t \sim \mathcal{N}(0, 1).$$

This corresponds to a white noise process.

### Model 3: Misspecified, Linear Trend

$$x_t = \alpha + \phi x_{t-1} + \gamma t + \varepsilon_t, \quad \alpha = -6,\ \phi = 0.6,\ \gamma = 0.06.$$

The series is standardized after generation.

### Model 4: Misspecified, Non-linear / Switching

$$x_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 0.5^2),$$

where $\mu_t$ switches deterministically between segments:

$$\mu_t = \begin{cases} -2, & t \in [1, 30] \cup [56, 95] \cup [106, 165], \\ 2, & t \in [31, 55] \cup [96, 105] \cup [166, 200]. \end{cases}$$

This generating mechanism could be seen as sampling from state-dependent Gaussian distributions, where observations at time points 1-30, 56-96, and 106-165 are generated by the distribution associated to state 1, and the remaining observations are generated by the distribution associated with state 2. This explains why this time series is modeled well with a Hidden Markov Model, which has precisely this structure (see Appendix F).

### Model 5: Misspecified, Heteroscedasticity

$$x_t = \alpha + \phi x_{t-1} + \varepsilon_t, \quad \alpha = 0,\ \phi = 0.6,$$

with time-varying innovation variance

$$\varepsilon_t \sim \mathcal{N}\left(0,\ \sigma_t^2\right), \quad \sigma_t = 0.1 + 0.012t.$$

### Model 6: Misspecified, Seasonality

$$x_t = \alpha + \beta d_t + \phi x_{t-1} + \varepsilon_t, \quad \alpha = -1,\ \phi = 0.6,\ \beta = 2,$$

where $d_t$ is a binary weekend indicator taking value 1 on weekend days ($t \equiv 6, 7 \pmod 7$) and 0 otherwise.

### Model 7: Misspecified, State-dependent Innovations

$$x_t = \alpha + \phi x_{t-1} + \varepsilon_t, \quad \alpha = 0,\ \phi = 0.6,$$

with state-dependent innovation variance

$$\varepsilon_t \sim \mathcal{N}\Big(0,\ \left(1 + \beta x_{t-1}\right)^2\Big), \quad \beta = 0.3.$$

### Model 8: Misspecified, Heavy-tailed Noise

$$x_t = \alpha + \phi x_{t-1} + \varepsilon_t, \quad \alpha = -1,\ \phi = 0.6,$$

with non-Gaussian innovations

$$\varepsilon_t \sim \text{Exp}(\lambda), \quad \lambda = 1.$$

That is, the noise follows an exponential distribution with mean 1 and variance 1, inducing heavy-tailed dynamics.

# C Deriving Prediction Errors for AR(1) Model

Here we derive the expected proportion of explained variance ($R^2$) and the Root Mean Squared Error (RMSE) from a given AR(1) model, assuming that the AR(1) model is correctly specified.

## C.1 Proportion of Variance Explained

$$\text{AR(1):} \quad X_t = \phi X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

The stationary variance of $X_t$ is

$$\text{Var}(X_t) = \frac{\sigma^2}{1 - \phi^2}.$$

The one–step predictor is $\hat{X}_t = \phi X_{t-1}$, with prediction error $\varepsilon_t$, so

$$\text{Var(error)} = \sigma^2.$$

Hence the proportion of variance explained is

$$R^2 \ = \ 1 - \frac{\text{Var(error)}}{\text{Var}(X_t)} = 1 - \frac{\sigma^2}{\sigma^2/(1-\phi^2)} = \phi^2.$$

For Model 1 in Appendix B this means that the expected $R^2 = \phi^2 = 0.6^2 = 0.36$.

## C.2 RMSE

$$\text{Var}(e_t) = \sigma^2, \quad \text{MSE} = \mathbb{E}[e_t^2] = \sigma^2,$$

and the root mean squared error is

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\sigma^2} = \sigma.$$

For Model 1 in Appendix B this means that the expected RMSE $= \sigma = 1$.

# D Three Additional Examples of Model Misspecification

In Section 3 we discussed three types of model misspecification which are perhaps the most common ones in psychological time series. Here we discuss three additional ones.

The first example, shown in the first row 8, is a time series that contains a seasonal component. Specifically, we simulated from an AR(1) model plus a weekend effect that increases the value of the variable during weekend. This effect is relatively hard to spot from the first plot showing the observed data and predictions. We might spot that the time series contains unusually regular spikes across time. The plot showing the residuals across time shows the same pattern. The plot showing residuals as a function of predictions does not add to the previous plots. The most useful plot is perhaps the final one showing the time series simulated from the estimated AR(1) model, which seems less regular than the observed time series. This type of misspecification is relatively hard to spot from the four diagnostic plots, despite the fact that the weekend effect we used to simulate the data was relatively large (see Appendix B). The smaller these effects are, the harder they will be to spot in diagnostic plots, which points towards the limits of what can be achieved by them (for more of a discussion of this point see Section 2.2.3).

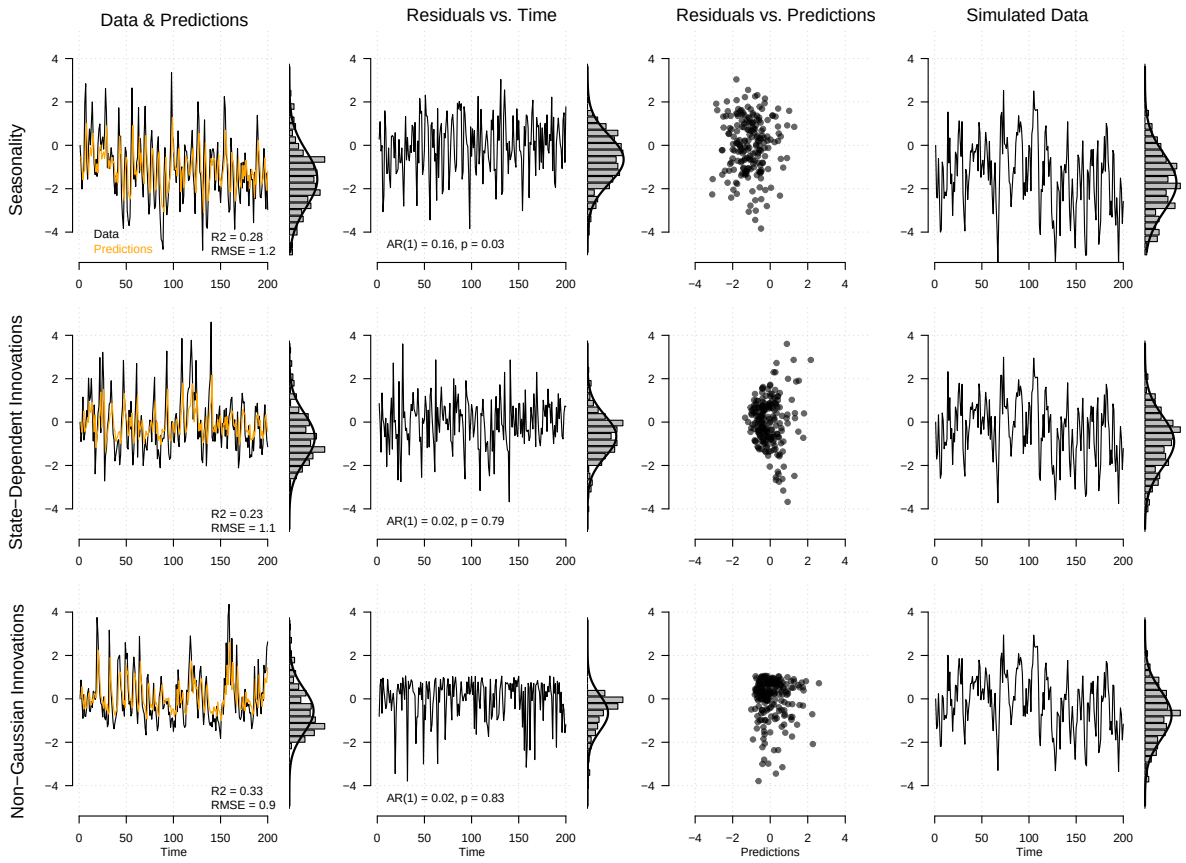


Figure 8: Three additional examples of misspecification. Row 1: Seasonality; Row 2: state-dependent innovations; and row 3: non-Gaussian innovations.

The second example, shown in the second row of Figure 8, is a time series in which the innovation variance is dependent on one of the modeled variables. Since in the AR(1) model there is only a single predictor, the variance depends on $X_{t-1}$ (see Appendix B). The first plot showing the observed time series looks close to what an AR(1) model can generate, except that the distribution seems a bit skewed, with more extreme positive than negative values. We see that the model does explain some variance ($R^2 = 0.23$), however, we have already seen that predictive performance is a poor indicator of model fit. The plot showing the residuals across time suggests that the residuals are uncorrelated and have the same, roughly Gaussian, distribution across time, therefore not indicating any misspecification. However, the third plot showing residuals as a function of the predictions clearly indicates model misspecification: we see that the variance of residuals seems to be a function of the value of the predicted value, with higher values being associated with higher residual variances. This makes sense, since the predictions are equal to $\phi X_{t-1}$ and the innovations are a function of $X_{t-1}$ (see Appendix B). Finally, the simulated data also suggests that there is some kind of model misspecification, because it does not show the skewness we see in the observational data.

The third example, shown in the last row of Figure 8, is an AR(1) model in which the innovations are not Gaussian distributed. In the first plot, we see that the observed data are skewed, which suggests that also the residuals are skewed. This is confirmed by the second plot, which shows us that the residuals are indeed heavily skewed, showing that the AR(1) model is misspecified. These types of time series are not uncommon in empirical time series. For example, for a variable like Anger, it often is the case that a person is largely not angry, but then if there is an anger-inducing event, there is an anger intensity depending on the event. Collapsed across time, this gives rise to distributions shown in this example. The third plot shows again the highly skewed distribution but does not add to the previous plot. Finally, the final plot showing the time series simulated from the fitted model does not show the skew, indicating model misfit.

Using intuitions from regression analysis, one could think of another type of misspecification, in which $X_t$ and $X_{t-1}$ are not linearly related like in the AR model, but in a non-linear way, such as:

$$X_t = \phi X_{t-1}^2 + \varepsilon_t$$

One would then expect to see this non-linearity in the plot showing residuals as a function of

predicted values. However, in the context of dynamic models, such models turn out to be either (a) diverging, a behavior that cannot be observed with the bounded scales used in psychology research, (b) bistable as shown in the first example in section 3, or (c) dampened by additional non-linear terms to the extent that the relationship is essentially linear again. We therefore did not include such a non-linear misspecification as a separate example.

# E Distribution of Prediction Errors Across Persons

Here we evaluate within-sample predictive performance quantified by the proportion of explained variance $R^2$ and the Root Mean Squared Error (RMSE) separately for each person and variable. Figure 9 shows the distribution of $R^2$ (left panel) across persons, separately for the four modeled variables, as violin plots, with the raw data plotted on top. The median $R^2s$ are 0.13 (Happy), 0.10 (Relaxed), 0.08 (Sad) and 0.04 (Angry) showing that for the typical person, we explain relatively little variance and we explain less variance for the two negatively valenced emotions. At the same time, we see that there is considerable variability across persons in how well the four emotions at time point $t$ are able to predict a given emotion at time point $t + 1$.

The right panel of Figure 9 shows the distribution of RMSEs across persons, separately for the four variables. The median RMSEs are 0.95 (Happy), 0.96 (Relaxed), 0.96 (Sad) and 0.99 (Angry). Since the data are scaled, on average, predicting with the mean would lead to a RMSE of 1. This shows that, the prediction performance for the typical person is relatively low. The two negatively valenced emotions are again a bit harder to predict. We also see again considerable variability across persons, but less than for $R^2$.

It is important to note that we computed all prediction errors within-sample. If we were to compute these prediction errors in an independent time series of the same individuals, we would expect that the $R^2$s would be smaller, and the RMSEs larger. If no independent time series is available, one could approximate the prediction errors in the population using a k-fold or rolling window cross-validation.

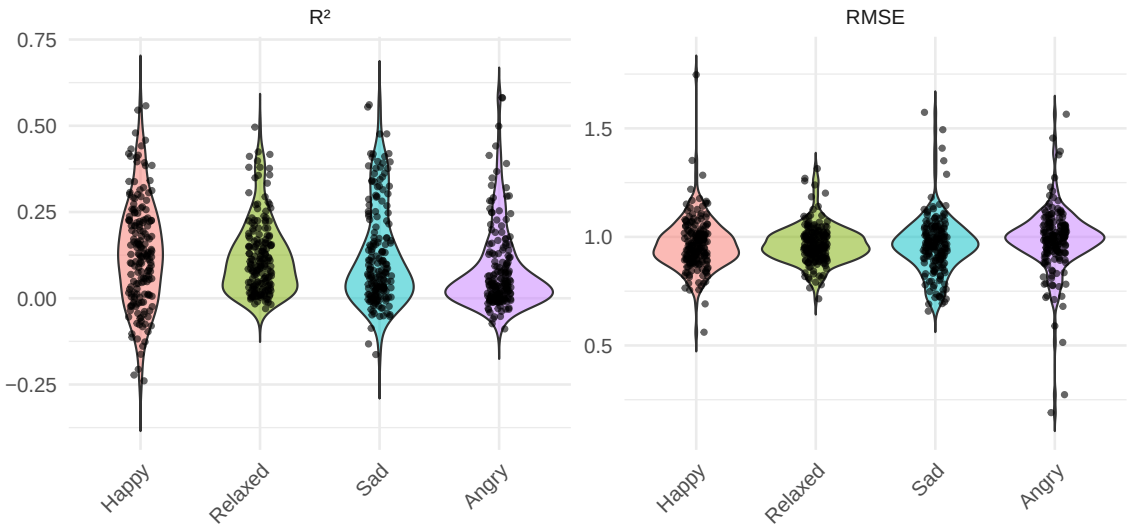


Figure 9: Left: Distribution of $R^2$ across persons summarized as a box plot, separately for the four modeled emotions; right: Distribution of the RMSE across persons summarized as a box plot, separately for the four modeled emotions.

The $R^2$ quantifies the proportion of variance in the outcome explained by the model, taking values between 0 (no explanatory power) and 1 (perfect fit). In contrast, RMSE is an absolute measure of prediction error, defined as the square root of the mean squared differences between observed and predicted values, and is expressed in the same units as the outcome. Thus, $R^2$ captures relative explanatory power, while RMSE captures the typical prediction error in absolute terms. The two measures of prediction error are typically moderately correlated in empirical data (see Figure 10); however they do not have to be. For example, bias does not affect $R^2$, but does affect RMSE.

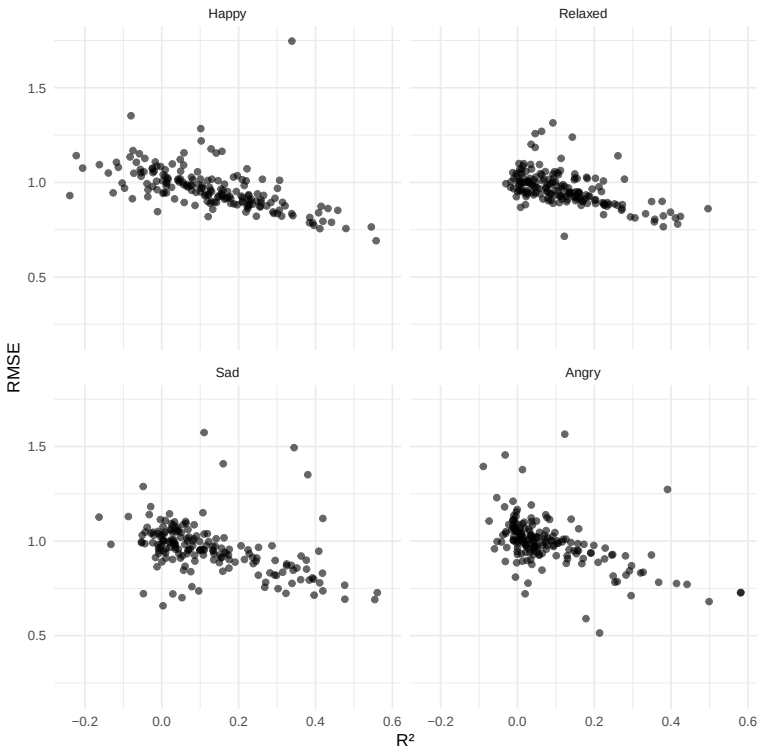


Figure 10: The relationship between $R^2$ and RMSE across persons, separately for the four modeled variables.

# F Assessing Model Fit of Hidden Markov Model

The first misspecification example shown in Figure 3 consists of a time series that is switching between discrete "states". In the main text, we discussed that such a time series could be, amongst other models, modeled well with a Hidden Markov Model (HMM). Here we explain this model in more detail, fit it to the same time series data as shown in the first row of Figure 3, and perform the same model checks, but this time for the HMM instead of the AR(1) model.

A HMM assumes that the observations are generated by an unobserved (latent) state process $\{s_t\}_{t=1}^T$, where $s_t \in \{1, \ldots, K\}$ for some number of states $K$. The state process is a first-order Markov chain:

$$\Pr(s_t \mid s_{t-1}, s_{t-2}, \ldots) = \Pr(s_t \mid s_{t-1})$$

meaning that the observation $x_t$ is independent from all other time points when conditioning on the state $s_t$ at time $t$

$$x_t \mid (s_t = k) \sim \mathcal{N}(\mu_k, \sigma_k^2),$$

with state-dependent mean $\mu_k$ and standard deviation $\sigma_k$.

We fitted this model using the `depmixS4` package (Visser & Speekenbrink, 2010) and specified two $K = 2$ Gaussian states to obtain maximum likelihood estimates of the parameters:

- initial state probabilities $\pi_k = \Pr(s_1 = k)$,
- transition probabilities $a_{ij} = \Pr(s_t = j \mid s_{t-1} = i)$,
- state-dependent emission parameters $(\mu_k, \sigma_k)$.

After estimating the HMM, we computed the posterior probability of each state at each time point:

$$\gamma_t(k) = \Pr(s_t = k \mid x_1, \ldots, x_T).$$

The expected value of $x_t$ under the model is then given by a weighted average of the state-dependent means:

$$\hat{x}_t = \sum_{k=1}^{K} \gamma_t(k)\, \mu_k.$$

We now have a prediction for each time point and can perform the same model checks as for the AR(1) model. We also simulated a time series of the same length from the model by first sampling a latent state sequence according to the estimated transition probability matrix and then sampling observations according to the corresponding state-dependent Gaussian distribution.

Figure 11 shows the same model checks we perform for the AR(1) model in the main text. The first plot shows that the model predictions follow the empirical data closely. The second plot shows the residuals. The assumptions of the HMM also imply Gaussian distributed residuals that do not change across time and are uncorrelated across time. We see that this is the case, suggesting that the HMM fits these data well. The residuals vs. prediction plot does not indicate any misfit. And the last plot displaying simulated data shows that the model generates data looking similar to the empirical data, which suggests that the model captures the empirical data well.

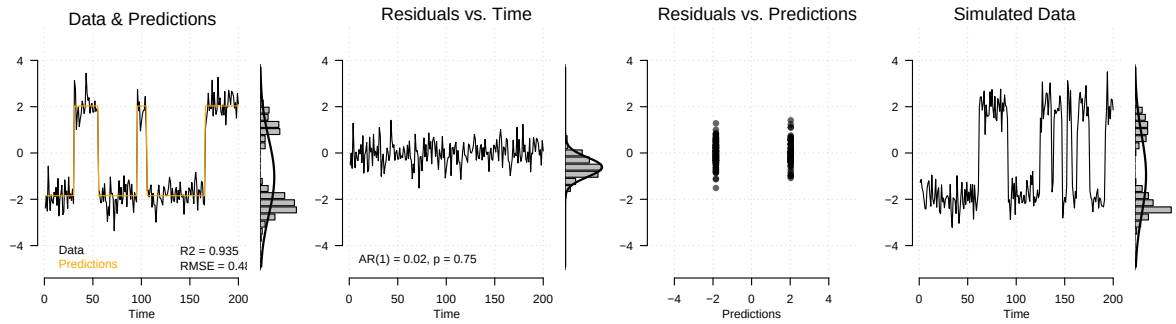


Figure 11: Model check diagnostics for the HMM estimated on the same time series as in the first row in Figure 3 in the main text.

This analysis shows that sometimes a model of a different family might be more appropriate and that the model checking tools presented in this paper for AR/VAR models can also be used for any other time series model.

# G Distribution of Random Effects of Lagged-Effects Parameters

Figure 12 shows the random effects distribution for the $4 \times 4 = 16$ lagged effects parameters of the VAR model estimated from the empirical data in Section 4. We see that most distributions are reasonably well approximated by a Gaussian distribution. Of course, the multilevel estimation shrinks parameters towards the grand mean. However, if the population heterogeneity would be clearly non-Gaussian, for example bimodal, then we would still see this in the distribution of the random effects estimates. Based on Figure 12, we would conclude that a multivariate Gaussian distribution is a reasonable model for the between-person heterogeneity in the lagged effects parameters, or at least that deviations from a multivariate Gaussian are too small to be reliably detected with the present sample size. We do not show the random effects distributions of the four intercepts, since we within-person centered the data, which means that those estimates have to be very close to zero.

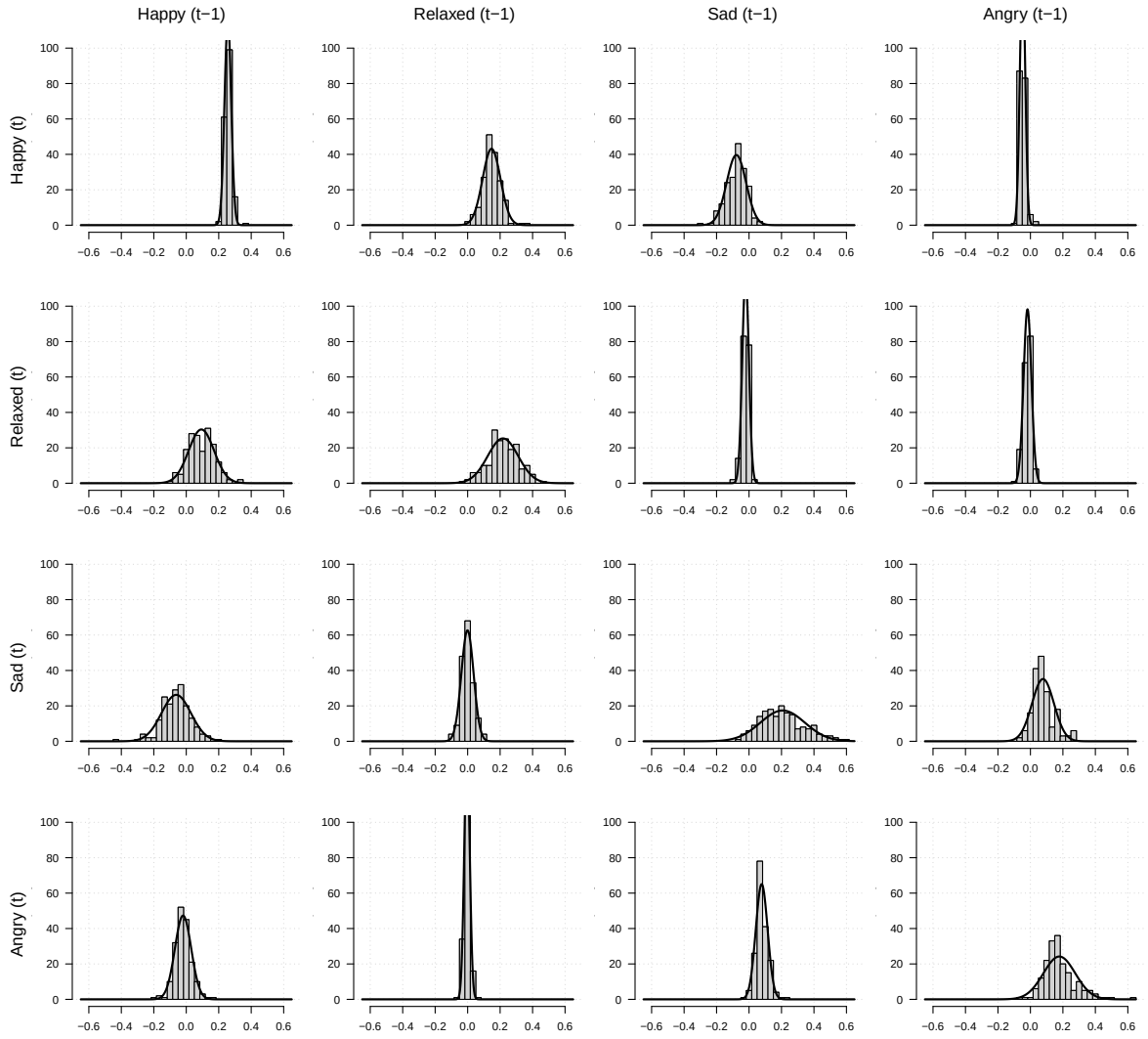


Figure 12: Distribution of the random effects estimates of the lagged effects for the 179 persons in the empirical dataset, displayed by histograms. The black line indicates the density of the best-fitting Gaussian distribution.

Figure 13 shows the random effects distributions of the intercepts. They are all centered at zero, since the package default of the *mlVAR* package is to grand-mean center the data. The large variation in intercepts for all variables shows that there are substantial between-person differences in how high they score on average in the four variables. Comparing the histograms with the best-fitting Gaussian distribution shows that the empirical distributions have heavier tails than a Gaussian, but the deviation is not large enough to clearly reject the Gaussian distribution as an approximation.

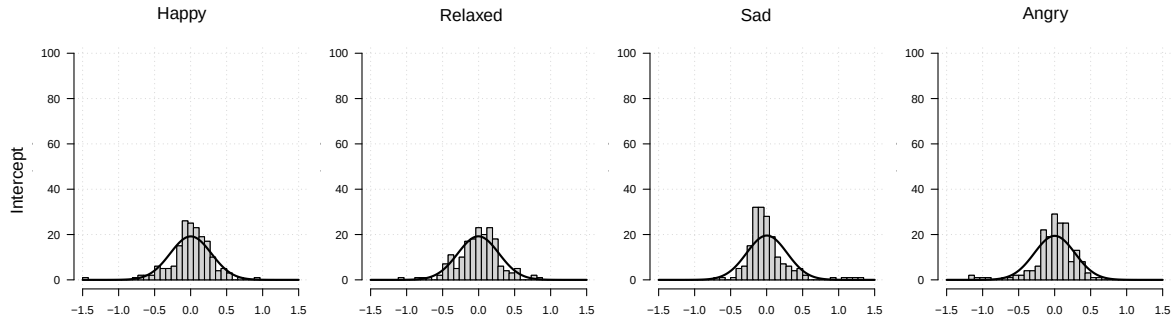


Figure 13: Distribution of the random effects estimates of the lagged effects for the 179 persons in the empirical dataset, displayed by histograms. The black line indicates the density of the best-fitting Gaussian distribution.