Behavior Research Methods (2024) 56:7152–7167
https://doi.org/10.3758/s13428-024-02413-4

ORIGINAL MANUSCRIPT

A new sample-size planning approach for person-specific VAR(1) studies: Predictive accuracy analysis

Jordan Revol[1] · Ginette Lafit[2] · Eva Ceulemans[1]

Accepted: 28 March 2024 / Published online: 8 May 2024
© The Psychonomic Society, Inc. 2024

Abstract

Researchers increasingly study short-term dynamic processes that evolve within single individuals using $N = 1$ studies. The processes of interest are typically captured by fitting a VAR(1) model to the resulting data. A crucial question is how to perform sample-size planning and thus decide on the number of measurement occasions that are needed. The most popular approach is to perform a power analysis, which focuses on detecting the effects of interest. We argue that performing sample-size planning based on out-of-sample predictive accuracy yields additional important information regarding potential overfitting of the model. Predictive accuracy quantifies how well the estimated VAR(1) model will allow predicting unseen data from the same individual. We propose a new simulation-based sample-size planning method called predictive accuracy analysis (PAA), and an associated Shiny app. This approach makes use of a novel predictive accuracy metric that accounts for the multivariate nature of the prediction problem. We showcase how the values of the different VAR(1) model parameters impact power and predictive accuracy-based sample-size recommendations using simulated data sets and real data applications. The range of recommended sample sizes is smaller for predictive accuracy analysis than for power analysis.

Keywords Predictive accuracy · Sample-size planning method · Power analysis · Monte Carlo simulation · Intensive longitudinal designs · Autoregressive models

Introduction

In psychological science, researchers increasingly study short-term dynamic processes that evolve within single individuals (Ariens et al., 2020; Hamaker & Wichers, 2017; Kuppens, 2015). To this end, they collect intensive longitudinal data (ILD) of one or more participants, using experimental paradigms (Vanhasbroeck et al., 2022) or experience sampling (Larson & Csikszentmihalyi, 2014; Myin-Germeys & Kuppens, 2021). Experience sampling implies that a participant is asked to report multiple times a day on momentary feelings, cognitions, and behaviors in daily life.

To capture one or more features of the dynamic process that underlies the ILD, time series methods are used. Although a multitude of methods has been proposed (for reviews see, Ariens et al., 2020; Hamaker et al., 2015; Vanhasbroeck et al., 2021), the most popular approach is the vector auto-regressive lag-one VAR(1) model. This model regresses all variable scores at a specific time point on the scores at the preceding time point (Lütkepohl, 2005). The resulting regression weights are then conceived to capture dynamic process features such as emotional inertia (Kuppens et al., 2010), and emotional augmentation and blunting (Krone et al., 2018; Kuppens et al., 2012; Pe et al., 2015). Moreover, building on the seminal work of Borsboom and Cramer (2013), many researchers have used the regression weights to construct dynamic networks of single participants (Fisher et al., 2017).

Despite the popularity of ILD and VAR(1) modeling, many open questions remain, a crucial one being how to perform sample-size planning. In the case of VAR(1), sample-size planning pertains to the number of time points that have to be collected for a specific participant (Trull & Ebner-Priemer, 2020). The importance of adequate sample-size planning for the replicability (Lakens, 2022: Munafó et al. 2017) and generalizability (see Trafimow, 2022) of the obtained VAR(1) results cannot be overstated (Kirtley, 2022;

✉ Jordan Revol
jordan.revol@kuleuven.be

[1] Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium

[2] Methodology of Educational Sciences Research Group, KU Leuven, Leuven, Belgium

Mansueto et al., 2022). Indeed, low sample size leads to high uncertainty of parameter estimates (i.e., large standard errors) and thus to low power. Moreover, it increases the risk of overfitting the data (see e.g., Bulteel et al., 2018a; Bulteel et al., 2018b; Yarkoni and Westfall, 2017). Overfitting (Babyak, 2004) occurs when the obtained VAR(1) parameters partly capture noise in the observed ILD. It is particularly prevalent when the number of variables increases and the sample size and the effect sizes drop (Hastie et al., 2013). As a result, the estimated parameters do not adequately reflect the underlying dynamic process and will not generalize well to unseen ILD data from the same participant. Hence, interpreting estimated parameters of overfitted models increases the risk of drawing erroneous conclusions about the dynamics of interest (Bulteel et al., 2018a).

The most popular approach to perform sample-size planning is to use power analysis (Button et al., 2013; Lakens, 2022). Power is defined as the probability of correctly rejecting the null hypothesis (e.g., a specific VAR(1) regression weight equals zero) when the alternative hypothesis is true (i.e., the VAR(1) regression weight differs from zero) (Cohen, 1992). The minimal sample size needed is then determined by checking whether the power is high enough. Although power analysis can sometimes be conducted using analytical formulas, this is not the case for VAR(1) based approaches because the formulas are lacking. Therefore, simulation-based approaches based on a Monte Carlo procedure (e.g., Green and MacLeod, 2016; Lane and Hennes, 2018) have been proposed and implemented in user-friendly Shiny applications (Lafit et al., 2021, 2022b).

These simulation-based approaches work as follows: first, the user specifies all the values needed to generate data according to the assumed underlying dynamic process (i.e., the VAR(1) parameters and the number of time points). Next, the user generates a large number of data sets using these values and fits a VAR(1) model to each of them. Finally, hypothesis tests are conducted for each VAR(1) parameter of interest and power is computed as the proportion of generated data sets in which the effect of interest is significant, given a predefined type I error rate. If the empirical power is too low, the sample size and thus the number of time points is increased until the empirical power is sufficiently high. The power calculations have to be done for each parameter of interest separately. Hence, when the statistical model includes more than one effect of interest, as is the case for the VAR(1) model, the minimal sample-size recommendation will almost always differ across these effects. This is illustrated in Fig. 1, where we obtain four different sample-size recommendations based on the specified VAR(1) parameters (parameters' values are presented in "The vector autoregressive model"). Such situations can be handled by selecting the largest recommended sample size across the four VAR(1) parameters.

Although hypothesis tests and power analysis have a long tradition in psychology, researchers can also take a different perspective on sample-size considerations and focus on overfitting and out-of-sample predictive accuracy performance rather than on testing effects of interest. In the person-specific VAR(1) setting this means that one is interested in the overall quality of the VAR(1) model, rather than in whether specific parameters deviate significantly from zero. One then wonders whether one has enough data to obtain a model that does not overfit the observed ILD and will generalize well to unseen data from the same participant. Overfitting might be a critical issue when fitting complex VAR(1) models to data with limited numbers of time points, which is sometimes the case in dynamical network analysis (Zhang et al., 2023). So far, the relation between sample size and predictive

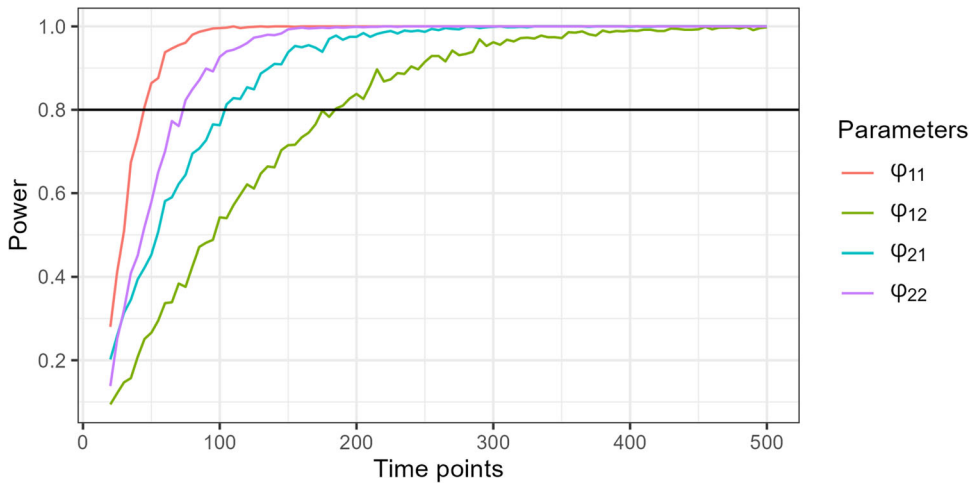


Fig. 1 Power as a function of the number of time points for the four regression weights of the VAR(1) model presented in "The vector autoregressive model"

accuracy is mainly studied from a post hoc perspective (i.e., after data collection) when trying to select the most optimal model from a set of possible models, often employing cross-validation methods (Liu & Zhou, 2023; Dejonckheere et al., 2019b; Lafit et al., 2022a). For the person-specific VAR(1) case, Bulteel et al. (2018a) and Lafit et al. (2022a) have shown that predictive accuracy of the VAR(1) model and thus also model selection depends on the sample size (i.e., the number of time points), the number of variables, and their correlations.

In this paper, we aim to use the predictive perspective to develop a new sample-size planning method for the person-specific VAR(1) case. This method, which we name Predictive Accuracy Analysis (PAA), determines the sample size needed to reach adequate predictive accuracy for a specific VAR(1) specification before collecting the data (i.e., a priori). We argue that, next to power analysis, PAA can provide valuable additional information when planning a study, through its focus on overfitting and the overall quality of the model. As mentioned above, overfitting is tightly related to the estimation precision of the parameters (Bulteel et al., 2018a), as high standard errors may be a sign of overfitting (Babyak, 2004). Hence, overfitting has negative consequences when conducting hypothesis testing. For the N=1 context, avoiding overfitting would thus lead to more trust in how well the model will predict future participant dynamics, but also in the reliability of the conclusions based on the hypothesis tests.

The remainder of this paper is organized as follows. In "The vector autoregressive model", we recapitulate the VAR(1) model and how to perform power analysis. Next, we introduce predictive accuracy analysis and the associated Shiny app. In "Simulation-based power analysis", we compare the sample-size recommendations obtained with power analysis and predictive accuracy analysis, by means of extensive simulation results in which we systematically manipulate the parameter values of the (V)AR(1) model and the sample size. In "Predictive accuracy analysis", we turn to a large database of time series of positive and negative affect and explore the range of recommended sample sizes. Finally, we conclude with a discussion.

The vector autoregressive model

The vector auto-regressive model of order one (VAR(1); Lütkepohl, 2005) is extremely popular in psychology to capture processes that take place in a single participant. In this model, each variable score y_{pt} ($p = 1, \ldots, P; t = 1, \ldots, T$) is linearly regressed on all P variables at the previous time point (i.e., $t - 1$), assuming that all time points are equally spaced:

$$\mathbf{y}_t = \boldsymbol{\delta} + \Phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \tag{1}$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{Pt})^T$ is a $(P \times 1)$ vector including the scores on the P variables at time point t. $\boldsymbol{\delta}$ is a $(P \times 1)$ vector of intercepts, and Φ is the $(P \times P)$ transition matrix, which includes the auto-regressive coefficients as diagonal elements and the cross-regressive coefficients as off-diagonal elements. Finally, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Pt})^T$ is the $(P \times 1)$ vector of innovations, which is assumed to be multivariate normally distributed with mean zero and covariance matrix Σ. These innovations reflect amongst other events that immediately impact the P variables. Through the transition matrix, this impact is also transmitted across time, a characteristic that differentiates innovations from measurement errors, which are not included in the VAR(1) model (Schuurman & Hamaker, 2019). For this paper, it is important to realize that innovations can and will usually be correlated, indicating that events impact the scores on multiple variables at once. The VAR(1) model implies a stationarity assumption, such that the variable means and (co)variances do not change across time. This assumption holds if all the absolute values of the eigenvalues of the transition matrix are smaller than one (Lütkepohl, 2005).

To further demonstrate the interpretation of the transition matrix and the variance–covariance matrix of the innovations, let us consider the following bivariate VAR(1) model:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 27 \\ 30 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.3 \\ 0.2 & 0.4 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{2}$$

with:

$$\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 4 \\ 4 & 10 \end{bmatrix} \right) \tag{3}$$

The auto-regressive and cross-regressive coefficients that range between .2 and .5 indicate how the variables exert influence on themselves and each other across time. In case the variables reflect self-reported emotional experience, the auto-regressive coefficients are usually positive and are assumed to quantify emotional inertia or the tendency for emotional experiences to linger. Positive cross-regressive coefficients suggest emotional augmentation where a higher score on one emotion predicts a higher score of another emotion at the next time point. Vice versa, negative cross-regressive coefficients would be interpreted as indicative of emotional blunting (Kuppens & Verduyn, 2017). Regarding the variance–covariance matrix, the innovations of both variables have a variance of 10 (translating into a standard deviation of approximately 3.16), indicating that quite some variance in the observed scores is due to external events, such as weather or interpersonal conflicts. Additionally, the innovations are contemporaneously related with a covariance of 4 (translating into a correlation of .4), implying that external events impact both variables to some extent.

The VAR(1) model includes a lot of parameters. Specifically, the number of VAR(1) parameters can be computed as $P + P^2 + (P + \frac{P^2-P}{2}) = \frac{3}{2}(P + P^2)$, and thus strongly depends on the number of variables. For example, in the case of two variables, the VAR(1) model contains nine parameters, as can also be observed in our example. In case only one variable is considered, the model reduces to the auto-regressive (AR(1)) model and implies three parameters: the auto-regressive effect, the intercept, and the innovation variance.

The VAR(1) model can be estimated in two different ways (Lafit et al., 2022a). First, the whole model can be estimated at once using dynamic structural equation modeling, implemented in the Mplus software (Hamaker et al., 2018), or using the lavaan R package (Rosseel, 2012). Second, the model can also be estimated using an ordinary least squares approach (OLS). This implies that we fit a standard regression model to each variable separately, each time using all the lagged variables as predictors. In this OLS approach, Σ must be estimated afterwards by calculating the variance–covariance matrix of the residuals of the separate regression models. In line with the predictive accuracy study of Lafit et al. (2022a), we rely on OLS because it provides closed-form solutions and can be straightforwardly implemented in open-source software. Computing predictive accuracy in such software is easy. Moreover, a small-scale simulation study showed that OLS performed equally well than Mplus and lavaan in terms of bias and precision of the different VAR(1) parameters (see supplementary materials). Note that the OLS framework also allows the implementation of significance tests on separate VAR(1) parameters. Researchers usually apply these tests on the auto- and cross-regressive coefficients (Ariens et al., 2020).

Simulation-based power analysis

Using the VAR(1) model example in "The vector autoregressive model", we will now illustrate how to conduct a simulation-based power analysis and perform sample-size planning to determine the minimal number of time points needed. This power analysis consists of four steps, which are shown in Fig. 2:

- Step 1. The researcher specifies the expected values of all VAR(1) model parameters (i.e., δ, Φ, Σ) and selects for which sample sizes power should be computed. As extensively described in recent papers on simulation-based power analysis (Lafit et al., 2021), the parameter values must be motivated based on literature or previous studies

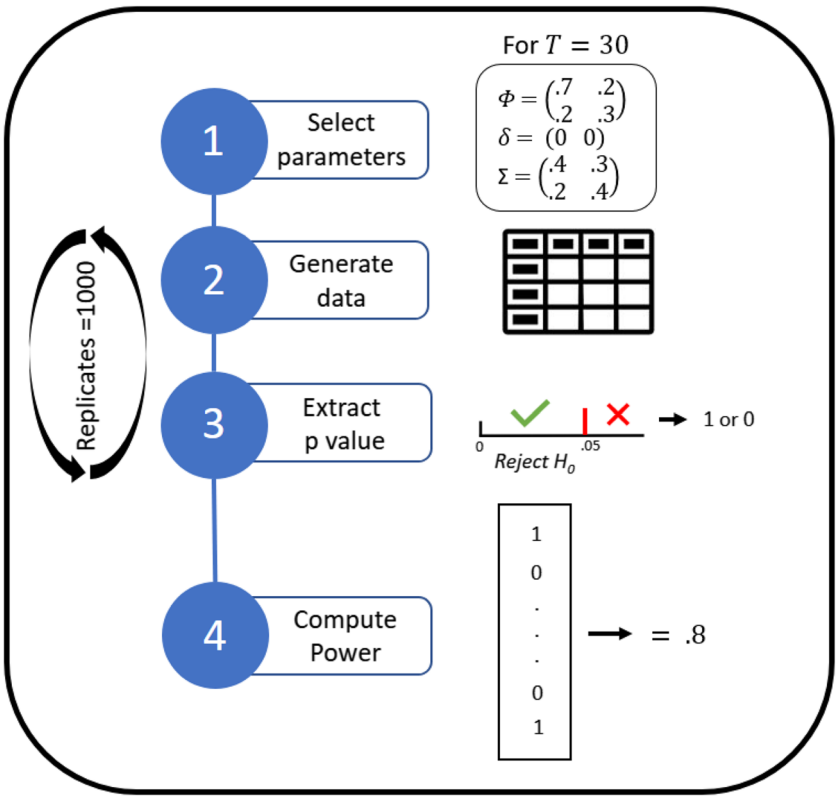


Fig. 2 Four steps of the simulation-based power analysis

(Lakens, 2022). We will use the values of our bivariate VAR(1) model example. As sample sizes (i.e., T), we selected values between 20 and 500 with steps of 5.

- Step 2. R datasets are generated based on the model and for each sample size that was specified in Step 1. The higher R, the more accurate the empirical power estimation will be. $R = 1000$ is usually deemed sufficient, though the empirical power estimate will still be subject to small sampling fluctuations.
- Step 3. A VAR(1) model is fit to each generated dataset and a t test is performed on each regression coefficient separately. The resulting p values (one for each t test) are compared to a chosen significance threshold (α), often being .05. It leads to reject H_0 if and only if $p < .05$.
- Step 4. Per regression coefficient and per sample size, the empirical power is given by the proportion of generated datasets for which H_0 was rejected. The results per regression coefficient can be visualized as a so-called power curve (see Fig. 1). Power is usually considered high enough when it is equal to or larger than .8. The minimal sample size needed is thus the lowest sample size for which the empirical power estimate exceeds .8. For our example, we can conclude that the lowest sample size needed ranges from 45 to 185 across the regression coefficients.

Predictive accuracy analysis

In this section, we introduce our novel predictive accuracy analysis (PAA) approach for sample-size determination, along with a new predictive accuracy metric. We will first present the four steps of PAA (Fig. 3), followed by a more detailed description and motivation for the new metric.

Four PAA steps

- Step 1. This step is identical to the first step in the power analysis and implies specifying the expected values of all VAR(1) model parameters and the to be considered sample-size values. For our example, we again use the parameter values of our bivariate VAR(1) example and select sample-size values between 20 and 500 time points using steps of 5.
- Step 2. Remember that PAA differs from power analysis in that it focuses on the generalizability of the estimated VAR(1) model to unseen data from the same participant. This means that we need to generate both training data, to estimate the model, and test data, to assess the model's generalizability to unseen data. Using the VAR(1) specifications in step 1, R training datasets are generated for each to be considered sample size, now labeled T^{Training}. Additionally, one test set is simulated, containing $T^{\text{Test}} = 100,000$ time points, and the associated $T^{\text{Test}} \times P$ innovation matrix $\mathbf{e}^{\text{True}}$ is stored. The length of T^{Test} has been established to ensure an acceptable computation time and a high accuracy in the estimation of predictive accuracy (see Step 3) for each replicate, which is crucial to obtain stable sample-size recommendations (see supplementary materials for more information).
- Step 3. A VAR(1) model is fit to each generated training dataset and used to predict the values of the test set, yielding a $(T^{\text{Test}} - 1) \times P$ prediction error matrix $\mathbf{e}^{\text{Pred},r}$. If a training data set yields model estimates that predict the test data well, the entries of $\mathbf{e}^{\text{Pred},r}$ should be close to those of $\mathbf{e}^{\text{True}}$. To capture the strength of this resemblance, we first compute the Mahalanobis distance between the prediction error vector of each time point and the true innovation distribution; the same distance transformation, which will be motivated in the next subsection, is executed for the innovations in $\mathbf{e}^{\text{True}}$. We then quantify the similarity between the two resulting vectors of distances by computing their squared Pearson correlation, R_r^2, which informs us about the proportion of variance in the distances for the true innovations that can be captured by the estimated model on the training data. Higher R_r^2 values reflect better predictive accuracy. Therefore, as demonstrated in Fig. 4, high enough T^{Training} values should yield higher R^2 values, whereas those for lower T^{Training} values will more often be smaller. Like in power analysis, where we compare the p value to a selected significance level, it is subsequently decided whether the obtained resemblance between the vectors of distances is strong enough, by checking if R_r^2 surpasses a targeted explained variance value R^2_{target} indicating satisfactory predictive accuracy.
- Step 4. The sufficient predictive accuracy probability (Sufficient PAP) is calculated as the empirical proportion of generated training datasets (associated with a T^{Training} value) for which the predictive accuracy is satisfactory (i.e., $R_r^2 > R^2_{\text{target}}$). In line with power analysis, we consider this proportion to be sufficiently high when it equals or exceeds 0.8. The minimum sample size required is thus the smallest T^{Training} value at which the Sufficient PAP surpasses 0.8. For our example, we can conclude that the lowest sample size needed amounts to 50 when setting R^2_{target} to 0.8, and to almost 100 when setting R^2_{target} to 0.9 (see Fig. 5). Hence, the selected value of R^2_{target} impacts the recommended sample size. In Fig. 6, we observe that as R^2_{target} increases, the recommended sample size T^{Training} increases rather linearly for $R^2_{\text{target}} < .9$, and grows faster afterwards. Therefore, balancing the recommended sample size and high generalizability, we

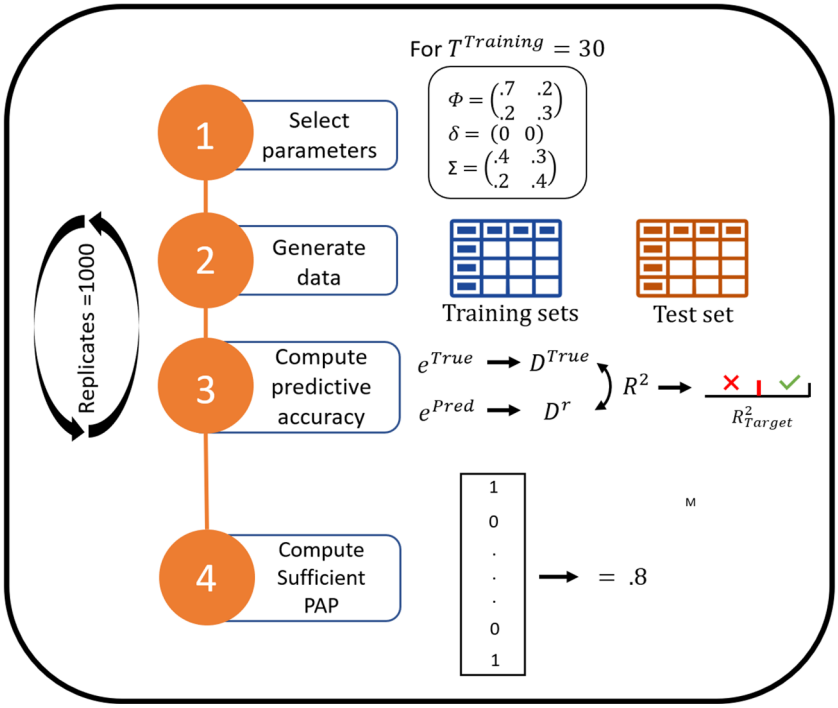


Fig. 3 Four steps of the simulation-based predictive accuracy analysis (PAA)

recommend setting R^2_{target} to .8 or .9, depending on how strict one wants to be.

Motivation for the new predictive accuracy metric

To motivate the new Mahalanobis-based predictive accuracy metric R^2, we elaborate on the issues of the mean squared prediction error (MSPE) metric that was used up till now in studies on VAR(1) predictive accuracy (Bulteel et al., 2018a, b; Lafit et al., 2022a). To introduce these issues, we first recapitulate the computation of the MSPE value. Remember that in Step 3, a VAR(1) model is fit to each r-th training set:

$$\mathbf{y}_t^{\text{Training},r} = \boldsymbol{\delta}^{\text{Training},r} + \Phi^{\text{Training},r}\mathbf{y}_{t-1}^{\text{Training},r} + \boldsymbol{\varepsilon}_t^{\text{Training},r} \quad (4)$$

where r ($r = 1, \ldots, R$) indicates the training set under consideration. Next, the resulting VAR(1) estimates are used to predict the test data using the r-th training data set:

$$\hat{\mathbf{y}}_t^{\text{Test},r} = \boldsymbol{\delta}^{\text{Training},r} + \Phi_p^{\text{Training},r}\mathbf{y}_{t-1}^{\text{Test}}, \quad (5)$$

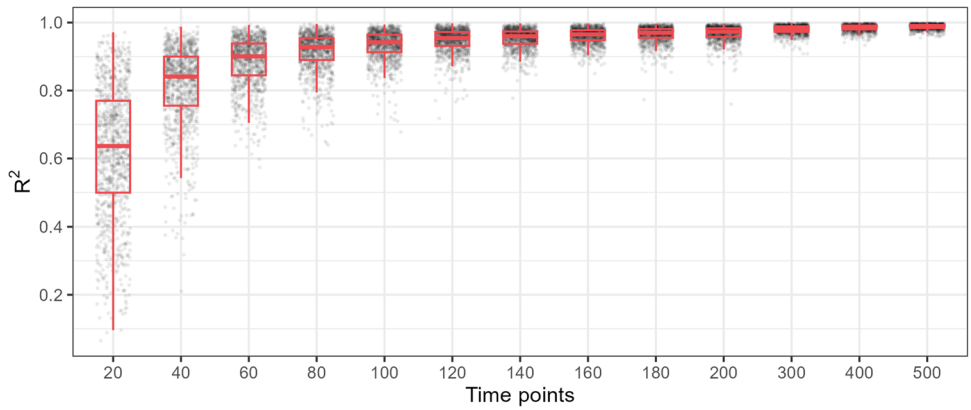


Fig. 4 Obtained distributions of R_r^2 values as a function of T^{training}

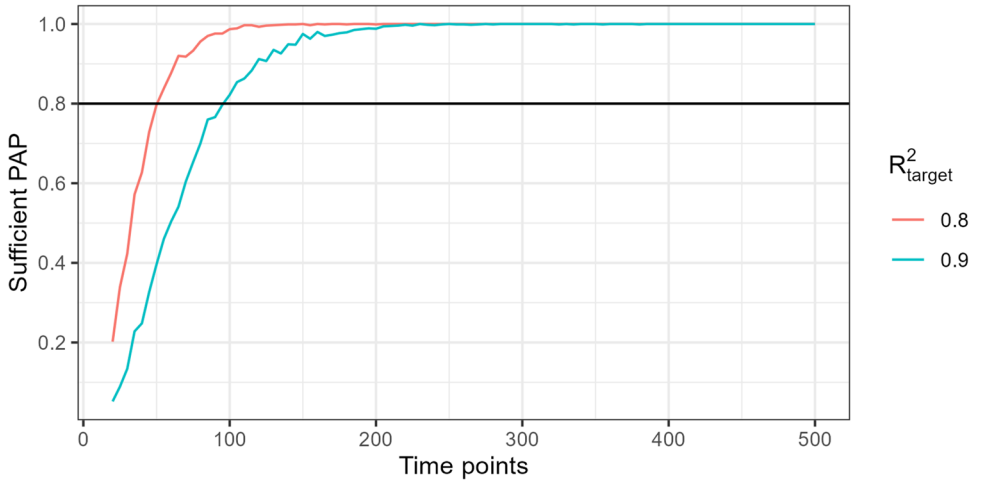


Fig. 5 Sufficient PAP curve for the VAR(1) example. The *black line* indicates the sufficient PAP target (i.e., .8) and the *colored lines* indicate different R^2_{target} values

The differences between the predicted values and the true ones are stored in the $((T^{test}-1)\times P)$ prediction error matrix $\mathbf{e}^{\text{Pred},r}$:

$$\mathbf{e}_t^{\text{Pred},r} = \mathbf{y}_t^{\text{Test}} - \hat{\mathbf{y}}_t^{\text{Test},r}. \tag{6}$$

Based on these prediction errors, the MSPE_t^r value for a given time point $t = (2, \ldots, T^{\text{Test}})$ equals

$$\text{MSPE}_t^r = \frac{1}{P}\sum_{p=1}^{P}(e_{p,t}^{\text{Pred},r})^2, \tag{7}$$

and the overall MSPE^r:

$$\text{MSPE}^r = \frac{1}{T^{\text{Test}}-1}\sum_{t=2}^{T^{\text{Test}}}\text{MSPE}_t^r \tag{8}$$

The intuition is that the lower the MSPE value, the better the predictive accuracy. For the AR(1) case, this makes sense because the obtained MPSE value will be bounded from below by the true innovation variance that was specified in Step 1, as shown by Bulteel et al. (2018a). For the VAR(1) model, however, interpreting the height of the MSPE value is more problematic because it does not properly account for the

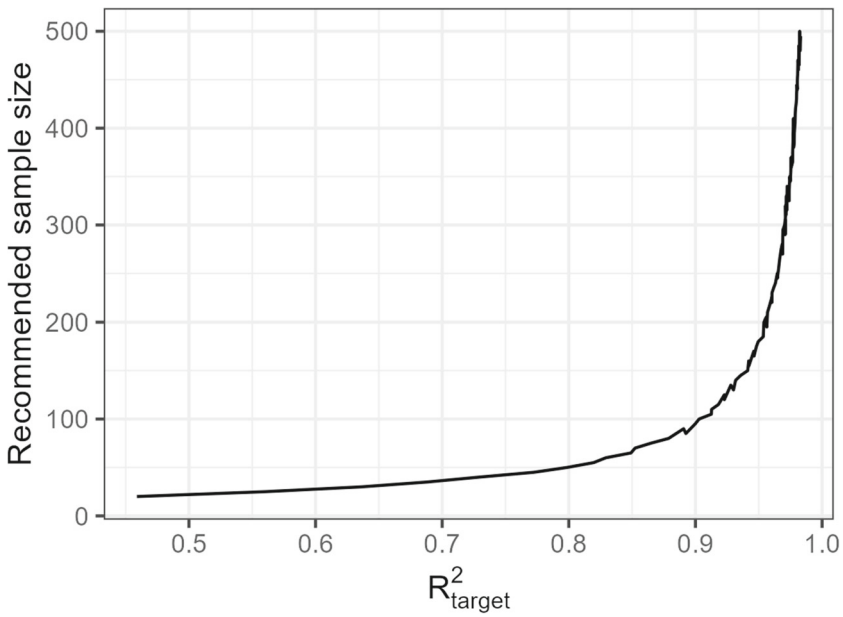


Fig. 6 Illustration of the exponential relationship between R^2_{target} and the recommended T^{Training}, as obtained for our VAR(1) example

multivariate nature of the prediction errors. It does not consider the covariance structure of the true innovations as well as potential between-variable differences in their variances. Figure 7 demonstrates this issue with MSPE. In the case of between-variable differences in innovation variance, the absolute squared prediction errors are expected to be smaller (resp. larger) for variables with a lower (resp. higher) innovation variance. These expectations are not accounted for when simply averaging the prediction errors across the variables. Indeed, having a very large prediction error on a variable with a small innovation variance can be compensated by having a small prediction error on a variable with a large innovation variance and yields the same MSPE_t as the opposite and more likely score pattern. In Fig. 7, this issue is highlighted by time points a and b having equal MSPE_t values, which does not reveal that a seems to be predicted worse than b. Moreover, in the case of strongly positive innovation covariances, we would expect that the prediction errors for a specific time point have the same sign, and thus, that time points for which the prediction errors have unequal signs should have a worse predictive accuracy value. However, this does not happen when computing the MSPE values, since squaring the prediction errors per variable removes the information about unexpected sign patterns. In Fig. 7, time points b and c have equal MSPE_t values, whereas the predictions for time point c clearly deviate more from the expected distribution.

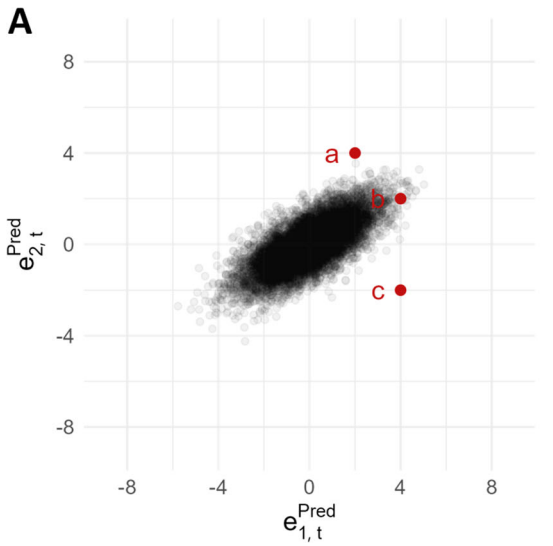
To avoid these drawbacks of the MSPE, we propose to use a different predictive accuracy metric that fully accounts for the innovation covariance matrix Σ. Building on the knowledge that the prediction errors would equal the innovations in case our trained model would be identical to the true model, we compute the Mahalanobis distance between the prediction error matrix $\mathbf{e}^{\text{Pred},r}$ and the multivariate innovation distribution.

$$D^r = (\mathbf{e}^{\text{Pred},r})^T \Sigma^{-1} \mathbf{e}^{\text{Pred},r}, \tag{9}$$

yielding a $((T^{\text{Test}} - 1) \times 1)$ vector of distances. From Fig. 7, we see that these distances capture the differences in predictive accuracy between time points a, b, and c. Next, replacing $\mathbf{e}^{\text{Pred},r}$ by $\mathbf{e}^{\text{True}}$ yields D^{True}. When T^{Training} is large enough, D^r should approximate D^{True}, whereas a too small T^{Training} value will yield more dissimilar distance vectors. Therefore, to quantify the predictive accuracy of the VAR(1) model estimates for the r-th training set, we compute the squared Pearson correlation R_r^2 between D^r and D^{True}.

Shiny app

We developed a user-friendly Shiny app (Chang et al., 2023) to compute the Sufficient PAP for specific T^{Training} values and compare it to the empirical power results. This Shiny app (see Fig. 8) is available on GitLab (https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0148925/shinyapp-paa_var_n1). Instructions to download and run the application locally are provided in the readme file of the repository.



B

t	$e^{Pred}_{1,t}$	$e^{Pred}_{2,t}$	$MSPE_t$	D
a	2	4	10	20
b	4	2	10	8
c	4	-2	10	40

Fig. 7 Plot A shows the expected distribution of prediction errors given $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$. Table B displays the prediction errors of the first variable (i.e., $e^{Pred}_{1,t}$) and the first variable (i.e., $e^{Pred}_{2,t}$) associated with the three time points a, b, c. The three time points give equal MSPE_t values but different D Mahalanobis distance scores



Fig. 8 Screenshot of the Shiny app PAA

Users first have to specify a seed number (i.e., for reproducibility purposes), the number of replicates (i.e., 1000 is recommended), and the number of measurement occasions (T^{Training}). Multiple values of T^{Training} can be provided by separating them with commas. To run PAA for our VAR(1) example, we can keep the default value for the number of replicates, and we can set "30, 60, 90, 120" as T^{Training} values. Second, users have to specify a R^2_{target} (in our case .9) and which PAP across the replicates they would deem sufficient (in our case .8). Third, to run the power analysis, the user selects the significance level α and the empirical power value that they would consider acceptable (e.g., .80). After indicating the number of variables for the VAR(1) model, the vector with intercepts $\boldsymbol{\delta}$, the transition matrix Φ, and the covariance matrix Σ have to be specified in a matrix. Generic names are given to the variables (i.e., Y_1, Y_2, etc.) but can be changed (e.g., PA, NA.). For our VAR(1) example, we specified two variables and entered $\boldsymbol{\delta}$, Φ, and Σ of the VAR(1) model example (see "The vector autoregressive model"). Note that decimal values should be indicated with a dot (e.g., .2, .5) and not a comma. Instructions are also included in the application by clicking the info buttons at the top right corner of each box.

The application displays the power analysis and PAA results for the considered T^{Training} values in two different tabs. Note that the empirical power curves are given for each of the VAR(1) parameters separately (each power curve can be highlighted by clicking the associated box). The application also creates a table indicating the sufficient PAP and the empirical power for each considered number of measurement occasions. A third tab allows to visualize simulated time series and to generate a dynamical network plot of the specified VAR(1) model. Finally, you can generate a PDF report of the PAA and the power analysis by clicking on "Generate report". It is important to note that the simulation can take a considerable amount of time depending on the number of T^{Training} values indicated. For instance, the simulation for our example took around 15 min using a Desktop PC with an 11th generation Intel Core i7KF processor (2 x 3.0 GHz) and 64 GB RAM.

Simulation studies

We ran three simulation studies to investigate how the predictive accuracy and power performance values and associated sample-size recommendations vary across different (V)AR(1) model specifications. Note that we focus on the power for the auto- and cross-regressive effects since these

parameters are almost always the only ones of interest. The first study focuses on the simpler AR(1) model, while the second and third studies investigate the VAR(1) case. Whereas the AR(1) and VAR(1) model specifications in the first two studies were selected by us building on earlier literature (Lafit et al., 2022a), they were derived from available empirical data in the third study to ensure the ecological validity of our findings. Rather than using the Shiny app, we ran the simulations in Julia (version 1.10.0; Bezanson et al., 2012)[1] to increase computation speed. For instance, the simulation to conduct the PAA for the example in "The vector autoregressive model" took 15 min in R and only 3 min and 30 s in Julia.

Study 1: AR(1) model

Simulation design

In the AR(1) study, we systematically varied the values of the three AR(1) parameters:

- the intercept δ was set to 0 or 5.
- the auto-regressive parameter ϕ value was set to .2, .5 or .8.
- the variance of the innovations σ was set to 1, 25, or 100.

Fully crossing the specification of these three parameters yielded 18 design cells. The set of T^{Training} values varied between 20 and 500 with steps of 5, implying that 97 values for T^{Training} were used. For every generated time series (i.e., training and testing datasets), we used a burn-in phase of 1000 time points[2] (Bulteel et al., 2018a). For the predictive accuracy analysis, we set the R^2_{target} to .8 and .9, and the sufficient PAP threshold to .80. We used 1000 replicates and T^{Test} equal to 100,000. For the power analysis, we used $\alpha = .05$ and .8 as empirical power target.

Results

Table 1 summarizes the minimal sample size recommendations based on PAA and power analysis as a function of the specified ϕ, δ, σ values. When the sample-size recommendation was higher than the largest considered T^{Training} value of 500, it was indicated by 500+.

Table 1 AR(1) simulations: Sample-size recommendations based on PAA, for $R^2_{target} = .9$ and $R^2_{target} = .8$, and on power

Design			Sample-size recommendations			
ϕ	δ	σ	$R^2_{target} = .9$	$R^2_{target} = .8$	Power ϕ	Power δ
0.2	0	1	65	30	215	–
0.2	5	1	70	30	215	20
0.2	0	25	65	30	215	–
0.2	5	25	65	30	215	20
0.2	0	100	70	30	215	–
0.2	5	100	65	30	215	40
0.5	0	1	65	35	40	–
0.5	5	1	70	35	40	20
0.5	0	25	70	35	40	–
0.5	5	25	70	35	40	25
0.5	0	100	70	35	40	–
0.5	5	100	65	35	40	45
0.8	0	1	95	55	20	–
0.8	5	1	95	55	20	45
0.8	0	25	95	55	20	–
0.8	5	25	95	55	20	50
0.8	0	100	95	55	20	–
0.8	5	100	95	55	20	70

In addition to the previously demonstrated effect of R^2_{target}, we can conclude that the sample-size recommendation for PAA clearly depends on the value of the auto-regressive parameter ϕ, with higher ϕ values being associated with more stringent sample-size recommendations. This result is probably due to the well-known small sample bias resulting in too low estimates of ϕ (Marriott & Pope, 1954) and thus in lower predictive accuracy. This small sample bias is known to exacerbate higher ϕ values (see supplementary materials for results on the bias of the estimates). Interestingly, this PAA result contrasts with that for power, where higher ϕ values are associated with lower sample-size recommendations. Moreover, the range of recommended sample sizes when manipulating ϕ seems to be larger for power analysis (i.e., between 20 and 220) than for PAA (i.e., between 30 and 55 for $R^2_{\text{target}} = .8$ and between 65 and 95 for $R^2_{\text{target}} = .9$).

Study 2: VAR(1) model

Design

Using the same PAA and power analysis settings, we now turn to bivariate VAR(1) models, implying that we also

[1] We used the following packages: DataFrames (version 1.6.1), DataTables (version 0.1.0), DataAPI (version 1.1.0), CSV (version 0.10.12) to handle the data; LinearAlgebra (version 0.5.1), GLM (version 1.9.0), HypothesisTests (version 0.11.0), StatsBase (version 0.34.2) to estimate the model and extract the estimated parameters; Distributions (version 0.25.107) and Distances (version 0.10.11) to handle statistical distributions.

[2] When generating (V)AR(1) time series, we have to use starting values, that is, the variable scores at the first time point. To remove the influence of these starting values, we removed the first 1000 time points (known as the burn-in phase).

Table 2 Specification of the transition matrix Φ (auto- and cross-regressive coefficients) for simulation study 2

Cross-regressive (CR)	Auto-regressive (AR)	
	Low	High
Low	$\begin{pmatrix} .4 & .1 \\ 0 & .3 \end{pmatrix}$	$\begin{pmatrix} .7 & .1 \\ 0 & .6 \end{pmatrix}$
High	$\begin{pmatrix} .4 & .3 \\ .2 & .3 \end{pmatrix}$	$\begin{pmatrix} .7 & .3 \\ .2 & .6 \end{pmatrix}$

include cross-regressive parameters and innovation covariances. Based on the AR(1) simulations where PAA was not impacted by δ, thus we decided to fix the intercepts to zero. As shown in Tables 2 and 3, we considered four different Φ specifications, manipulating the height of the auto-regressive (i.e., low and high) and cross-regressive (i.e., low and high) effects. Additionally, four different Σ specifications were used, manipulating the values of the innovation variances (same or not) and covariance (i.e., boiling down to a correlation of zero or of .6). Crossing the four Φ and Σ settings, 16 design cells were obtained.

Results

Table 4 summarizes the minimal sample-size recommendations based on PAA and power analysis for the 16 design cells.

Again, we see that the Φ values determine the PAA-based sample sizes, with higher auto- and cross-regressive effects leading to larger required sample sizes; the Σ specification does not play a role. The required sample sizes are higher than they were for AR(1), which is due to the increased model complexity. Moreover, power analysis again leads to a larger range of desired sample sizes than PAA, which are inversely related to the strength of the effects in question.

Study 3: Model specification based on empirical data

Data

Whereas the previous two simulation studies allowed us to systematically investigate the influence of each (V)AR(1) parameter on sample-size recommendations, we do not know if the studied model specifications are realistic. Hence, following (Loossens et al., 2021), we now investigate the sample-size recommendations that result from using the VAR(1) parameters obtained when analyzing empirical data as input for step 1 of the PAA analysis. Specifically, we re-analyzed the data of seven studies asking participants to report on their momentary positive (e.g., relaxed, happy, cheerful) and negative (e.g., angry, anxious, depressed, sad, stressed) affect multiple times a day for a few days using their smartphone (for more details see Table 5). Positive affect (PA) and negative affect (NA) variables were obtained for each participant by averaging the positive and negative items, respectively.

Then, a VAR(1) model was fit for each participant and the estimated parameters were used in the PAA and power analyses. Note that we only considered the 825 participants, for which the estimated Φ matrices had absolute eigenvalues smaller than one (Lütkepohl, 2005), and who filled in at least 75% of the momentary questionnaires. To handle night breaks in ESM data, missing observations were included at each end of the day. The estimation procedure followed a list-wise deletion procedure, meaning that observations with outcomes or predictors containing missing values were deleted.

Table 3 Specification of the innovation covariance matrix Σ for simulation study 2

Correlation	Variance	
	Equal	Not equal
.0	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & .0 \\ 0 & 25 \end{pmatrix}$
.6	$\begin{pmatrix} 1 & .6 \\ .6 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 3 \\ 3 & 25 \end{pmatrix}$

Results

Figures 9 and 10 display the recommended sample sizes resulting from predictive accuracy analysis and power analysis, respectively. Note that for the power analysis, we did not include cases where the effects of interest were smaller than .1. We again see that the range of sample sizes is much larger for power analysis than for predictive accuracy analysis and that the latter ones vary between 40 and 80 (resp. 85 and 135).

Discussion

In this paper, we developed a sample-size planning method, called predictive accuracy analysis (PAA), which allows optimizing the predictive accuracy of a to-be-estimated VAR(1) model toward unseen data from the same participant. Additionally, we introduced a new predictive accuracy metric that adequately accounts for the covariance matrix of the innovations. We also provide a Shiny app to run predictive accuracy analysis and compare the resulting recommended number of measurement occasions to that provided by power analysis.

We ran three simulation studies, to investigate how the results from both sample-size planning methods differ and how these results vary across VAR(1) specifications. Regard-

Table 4 VAR(1) simulations: Sample-size recommendations based on PAA, for $R^2_{target} = .9$ and $R^2_{target} = .8$, and on power

Design				Sample-size recommendations					
AR	CR	Variance	Covariance	$R^2_{target} = .9$	$R^2_{target} = .8$	Power ϕ_{11}	Power ϕ_{12}	Power ϕ_{21}	Power ϕ_{22}
Low	Low	Symmetric	Low	95	50	60	500+	500+	100
Low	Low	Symmetric	High	95	50	85	500+	500+	150
Low	Low	Asymmetric	Low	95	50	60	500+	500+	100
Low	Low	Asymmetric	High	95	50	85	500+	500+	150
Low	High	Symmetric	Low	100	50	60	95	165	95
Low	High	Symmetric	High	100	50	85	145	265	155
Low	High	Asymmetric	Low	100	50	65	500+	20	50
Low	High	Asymmetric	High	100	50	85	500+	20	70
High	Low	Symmetric	Low	110	60	30	500+	500+	35
High	Low	Symmetric	High	105	60	35	500+	500+	45
High	Low	Asymmetric	Low	105	60	30	500+	500+	35
High	Low	Asymmetric	High	105	60	35	500+	500+	45
High	High	Symmetric	Low	130	75	25	70	105	35
High	High	Symmetric	High	130	75	35	125	175	50
High	High	Asymmetric	Low	125	75	30	380	20	20
High	High	Asymmetric	High	125	75	35	500+	20	20

Table 5 Summary of the datasets used in simulation study 3

Data set	Participants	Compliance	Standard deviation compliance	Days	Observations per day	Number of parameters
Dejonckheere et al. (2019a)	100	0.888	0.096	14	7	90
Heininga et al. (2019)	118	0.871	0.111	7	10	101
Pe et al. (2016)	200	0.873	0.091	7	10	180
Pe et al. (2016)	190	0.879	0.090	7	10	171
Pe et al. (2016)	177	0.884	0.087	7	10	165
Provenzano et al. (2021)	34	0.874	0.104	7	10	28
Sels et al. (2017)	100	0.801	0.068	8	10	90

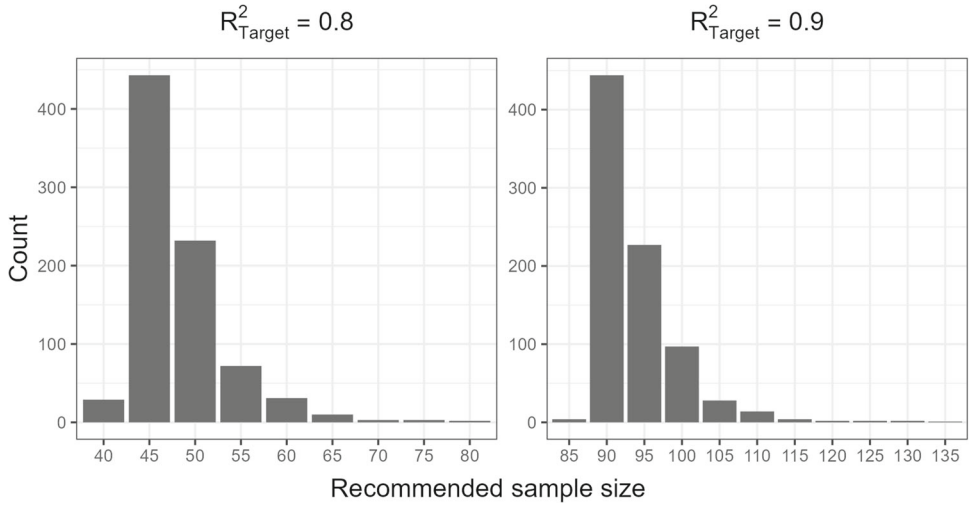


Fig. 9 Sample-size recommendations based on PAA using the empirical data sets

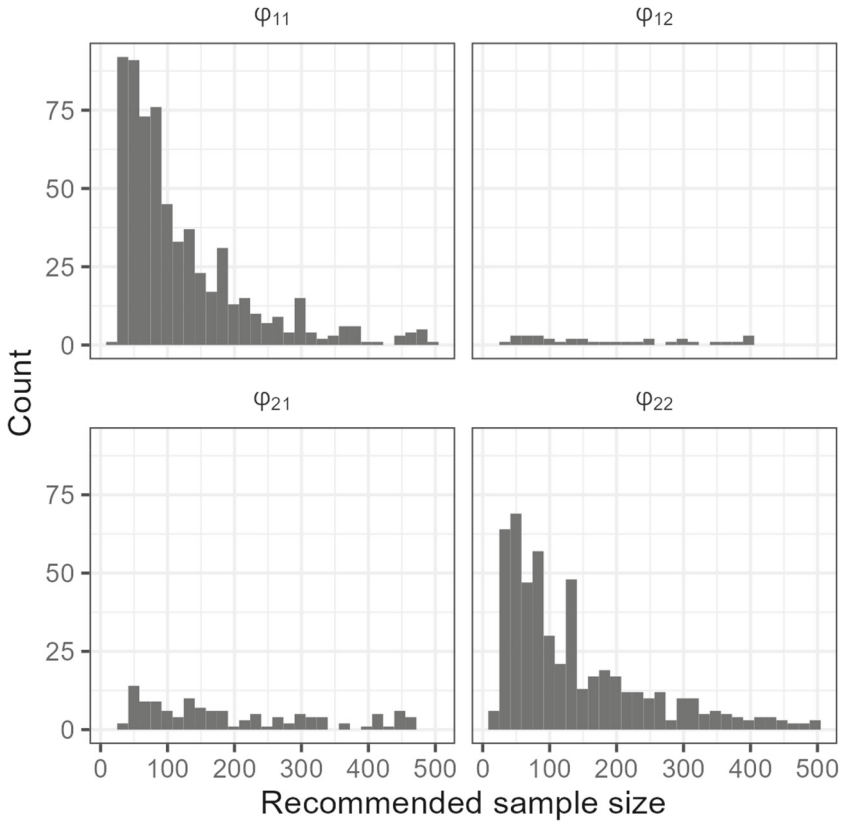


Fig. 10 Sample-size recommendations for the auto- and cross-regressive effects based on power using the empirical data sets

ing dependence on VAR(1) specifications, higher expected auto- and cross-regressive coefficients lead to higher recommended sample sizes, due to small sample bias in estimating them (see supplementary materials). Interestingly, the auto- and cross-regressive parameters have the opposite effect in power analysis: the higher their values, the lower the recommended sample size. In contrast, the innovation (co)variances do not seem to matter much for the recommended sample size, which makes sense given that we developed a new predictive accuracy metric that accounts for them. Over the three studies, we showed that the range of recommended sample sizes is lower for the predictive accuracy analysis than for power analysis. This is important and promising information since it implies that the recommended sample size based on PAA is less dependent on specific model parameter values and, hence, will be less impacted by uncertainty and misspecification of the VAR(1) model in step 1 of the analysis. Indeed, in sample-size planning research, how to obtain these parameter values is a crucial bottleneck and always comes with uncertainty (Lafit et al., 2021, 2023; Lakens, 2022; Lane & Hennes, 2018). Going a step further, guidelines can be derived. For instance, Study 3 shows that when modeling the dynamical interplay of positive and negative affect using a bivariate VAR(1) model, collecting 80 (resp. 135) time points would probably be sufficient, based on a R^2_{Target} of .8 (resp. .9) and a Sufficient PAP of .8.

As mentioned in the Introduction, we argue that PAA analysis complements power analysis and yields additional important information. Even when researchers are mainly interested in the significance of specific VAR(1) effects, evaluating whether the sample size suggested by the power analysis yields a high Sufficient PAP can nuance one's evaluation of a study design, in that it sheds light on the overall risk of overfitting. In case the PAA recommended sample size is lower than that of the power analysis, one's trust in the design is strengthened further. In the opposite case, it should spark caution, in that researchers should be aware that interpreting model features, other than the significance of the effect of interest, is not without risk. Stated differently, using PAA may avoid premature conclusions arising from potentially overfitting models, leading to unreproducible results (Yarkoni & Westfall, 2017).

Importantly, users should keep in mind that PAA is an a priori analysis that is conducted before collecting the data. Hence, conclusions on the predictive accuracy of the model that is fitted after data collection cannot be drawn from this analysis. Indeed, PAA builds on the premise that the specified VAR(1) parameters are correct and that the VAR(1) assumptions hold. If one or more of these assumptions do not hold,

actual predictive accuracy may be quite worse. Moreover, even if all assumptions are correct, Step 4 of PAA implies that we expect lower predictive accuracy in 20% of the cases. Therefore, we also recommend investigating the predictive performance of the estimated model on the collected data itself (i.e., post hoc), using a cross-validation procedure or a separate data set from the same participant. How predictive accuracy should be computed in this post hoc scenario and how large this separate data set should be, should be investigated further.

Directions for future research

We see several directions for future research. First, in the current simulation studies, we only investigated settings with one or two variables. Yet, in psychopathology research, sparked by the popularity of dynamic network modeling (Epskamp et al., 2018; Mansueto et al., 2022), researchers fit VAR(1) models to datasets with far more variables and relatively low numbers of measurement occasions. Since we showed that increasing the number of variables from one to two already leads to higher sample-size recommendations, more work is needed to investigate the consequences of our findings for dynamic network research and warn researchers and clinicians about potential generalizability issues.

Second, similar to power analysis (i.e., choice of α and desired level of power), predictive accuracy analysis requires specifying some analysis settings. Specifically, the researcher has to set the R^2_{target} that is applied per replicate to decide whether that replicate yielded estimates that lead to adequate prediction. Moreover, researchers have to set the Sufficient PAP threshold, which indicates which proportion of the replicates have to reach the R^2_{target}. The values that are selected obviously determine the recommended sample size, with higher values leading to higher recommended numbers of time points (see for instance Table 4). Just as one can debate whether a power of .8 or a type I error .05 is always adequate, further research into which PAA thresholds are meaningful for specific applications is warranted.

Third, we focused in this paper on the person-specific VAR(1) model. Yet, the proposed ideas can in principle be applied to extensions of the person-specific VAR model as long as these models allow us to decide for each measurement occasion in the test set which model parameters should be used to compute the predicted values. Examples of such models are the person-specific Threshold-VAR and the Fixed Moderated VAR models. The threshold VAR(1) model assumes there are two or more regimes that are governed by different VAR(1) parameters, and that regime-switching takes place based on the value of an observed variable (Hamaker et al., 2009; Tong & Lim, 1980). The Fixed Moderated VAR(1) model allows the size of all the VAR(1) parameters to depend on an observed covariate, to capture how they for instance are impacted by context information (Adolf et al., 2017; Phillips, 1995). Such extensions would benefit applied users a lot as they accommodate problems that are often encountered in real life, such as discrete parameter changes (e.g., after an intervention) or initial elevation (e.g., it may take some time before you observe a rather stable trend). Note however that information on when the intervention takes place or when the data stabilizes is crucial. Without such information, it is hard to mimic such scenarios when simulating data and to model the data appropriately. Another example is the continuous-time VAR model (De Haan-Rietdijk et al., 2017) which accommodates unequal time intervals between observations. For this model, the most important challenge is to make sure that the data in the training and test set are simulated in such a way that the distribution of the time intervals resembles the one that can be expected in the to be collected data. For other popular models, the extension is less straightforward and should be investigated further. A prime example is the multilevel extension of VAR(1) (Hamaker et al., 2018; Jongerling et al., 2015). In this case, PAA analysis should accommodate the random effects. Moreover, one should define to which participants one wants to generalize the estimated model, which can be the participants from the sample used to train the data, and/or new participants from the same population.

A fourth direction for future research concerns the above-mentioned impact of small sample bias on the PAA results. It would be worthwhile to investigate ways to correct this bias and implement well-performing approaches in PAA analysis. Indeed, to our knowledge, psychology research has simply ignored this phenomenon.

Conclusion

We introduced a new a priori sample-size planning method for VAR(1) analysis that capitalizes on the generalizability of the estimated model for a participant to unseen data from the same participant. We also provide a Shiny app that allows to run the analysis given expected VAR(1) parameters and compare the recommended sample size provided by power analysis. Through simulation studies, we showed that the recommended sample size varies less across model specifications than that of power analysis.

Supplementary Material

For reproducibility purposes, the files to conduct the simulation studies and supplementary material are available on GitLab (https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0148925/paa_var_n1). The Shiny application can be down-

loaded from a GitLab repository (https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0148925/shinyapp-paa_var_n1).

Additionally, results from three supplementary simulations can be found in the supplementary materials. First, to be certain that the results are not due to the seed number we chose for the main simulation, we reproduced the analysis with a new seed number which yielded equivalent results. Second, we showcased how the number of time points in the test set (i.e., T^{Test}) influences the estimation of the predictive accuracy. Finally, we conducted a concise comparative simulation to assess the performance of three distinct approaches: ordinary least squares (OLS), Mplus, and lavaan.

Author Contributions The authors made the following contributions. Jordan Revol: Conceptualization, Formal Analysis, Methodology, Visualization, Software, Writing - Original Draft Preparation, Review & Editing; Ginette Lafit: Conceptualization, Methodology, Supervision, Writing - Original Draft Preparation, Review & Editing. Eva Ceulemans: Conceptualization, Methodology, Funding acquisition, Supervision, Writing - Original Draft Preparation, Review & Editing.

Funding The research presented in this article was supported by research grants from the Fund for Scientific Research-Flanders (FWO; Project No. G0C9821N) and from the Research Council of KU Leuven (C14/23/062; iBOF/21/090) awarded to E. Ceulemans.

Declarations

Conflict of Interest The authors declare that there are no conflicts of interest with respect to the authorship or the publication of this article.

References

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.

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psycho-somatic Research, 137*, 110191.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421.

Bezanson, J., Karpinski, S., Shah, V., & Edelman, A. (2012). Julia: A fast dynamic language for technical computing.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*(1), 91–121.

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*, 740–756.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2018). Improved insight into and prediction of network dynamics by combining VAR and dimension reduction. *Multivariate Behavioral Research, 53*(6), 853–875.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafó, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376.

Chang, W., Cheng, J., Allaire, JJ., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). Shiny: Web application framework for R.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98–101.

De Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discretevs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology, 8*, 1849.

Dejonckheere, E., Kalokerinos, E. K., Bastian, B., & Kuppens, P. (2019). Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition and Emotion, 33*(5), 1076–1083.

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour, 3*(5), 478–491.

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

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology, 126*(8), 1044–1056.

Green, P., & MacLeod, C. J. (2016). SIMR?: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498.

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

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

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.

Hamaker, E. L., Zhang, Z., & Van Der Maas, H. L. J. (2009). Using threshold autoregressive models to study dyadic interactions. *Psychometrika, 74*(4), 727.

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., & Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: Positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry, 19*(1), 59.

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research, 50*(3), 334–349.

Kirtley, O. J. (2022). Advancing credibility in longitudinal research by implementing open science practices: Opportunities, practical examples, and challenges. *Infant and Child Development, 31*(1).

Krone, T., Albers, C. J., Kuppens, P., & Timmerman, M. E. (2018). A multivariate statistical model for emotion dynamics. *Emotion, 18*, 739–754.

Kuppens, P. (2015). It's about time: A special section on affect dynamics. *Emotion Review, 7*(4), 297–300.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science, 21*(7), 984–991.

Kuppens, P., Champagne, D., & Tuerlinckx, F. (2012). The dynamic interplay between appraisal and core affect in daily life. *Frontiers in Psychology, 3*.

Kuppens, P., & Verduyn, P. (2017). *Emotion dynamics. Current Opinion in Psychology, 17*, 22–26.

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science, 4*(1), 251524592097873.

Lafit, G., Meers, K., & Ceulemans, E. (2022). A systematic study into the factors that affect the predictive accuracy of multilevel VAR(1) models. *Psychometrika, 87*(2), 432–476.

Lafit, G., Revol, J., Cloos, L., Kuppens, P., & Ceulemans, E. (2023). *The effect of different operationalizations of affect and preprocessing choices on power-based sample size recommendations in intensive longitudinal research*.

Lafit, G., Sels, L., Adolf, J. K., Loeys, T., & Ceulemans, E. (2022b). PowerLAPIM: An application to conduct power analysis for linear and quadratic longitudinal actor–partner interdependence models in intensive longitudinal dyadic designs. *Journal of Social and Personal Relationships*, page 02654075221080128.

Lakens, D. (2022). Sample size justification. Collabra. *Psychology, 8*(1), 33267.

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships, 35*(1), 7–31.

Larson, R. & Csikszentmihalyi, M. (2014). The Experience Sampling Method, pages 21–34. Springer Netherlands, Dordrecht.

Liu, S. & Zhou, D. J. (2023). Using cross-validation methods to select time series models: Promises and pitfalls. *British Journal of Mathematical and Statistical Psychology*, page bmsp.12330.

Loossens, T., Dejonckheere, E., Tuerlinckx, F., & Verdonck, S. (2021). Informing VAR(1) with qualitative dynamical features improves predictive accuracy. *Psychological Methods, 26*(6), 635–659.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin Heidelberg: Springer.

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

Marriott, F. H. C., & Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika, 41*(3/4), 390.

Munafó, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 0021.

Myin-Germeys, I., & Kuppens, P. (Eds.). (2021). *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies*. Leuven: Center for Research on Experience Sampling and Ambulatory Methods.

Pe, M. L., Brose, A., Gotlib, I. H., & Kuppens, P. (2016). Affective updating ability and stressful events interact to prospectively predict increases in depressive symptoms over time. *Emotion, 16*(1), 73–82.

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

Phillips, P. C. B. (1995). Fully modified least squares and vector autoregression. *Econo-metrica, 63*(5), 1023.

Provenzano, J., Fossati, P., Dejonckheere, E., Verduyn, P., & Kuppens, P. (2021). In exibly sustained negative affect and rumination independently link default mode network efficiency to subclinical depressive symptoms. *Journal of Affective Disorders, 293*, 347–354.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91.

Sels, L., Ceulemans, E., & Kuppens, P. (2017). Partner-expected affect: How you feel now is predicted by how your partner thought you felt before. *Emotion, 17*(7), 1066–1077.

Tong, H., & Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological), 42*(3), 245–268.

Trafimow, D. (2022). Generalizing across auxiliary, statistical, and inferential assumptions. *Journal for the Theory of Social Behaviour, 52*(1), 37–48.

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology, 129*(1), 56–63.

Vanhasbroeck, N., Ariens, S., Tuerlinckx, F., & Loossens, T. (2021). Computational Models for Affect Dynamics. In C. E. Waugh & P. Kuppens (Eds.), *Affect Dynamics* (pp. 213–260). Cham: Springer International Publishing.

Vanhasbroeck, N., Loossens, T., Anarat, N., Ariens, S., Vanpaemel, W., Moors, A., & Tuerlinckx, F. (2022). Stimulus-driven affective change: Evaluating computational models of affect dynamics in conjunction with input. *Affective Science, 3*(3), 559–576.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.

Zhang, Y., Revol, J., Lafit, G., Ernst, A., Razum, J., Ceulemans, E., & Bringmann, L. (2023). Sample size optimization for person-specific temporal networks using power analysis and predictive accuracy analysis. Manuscript in preparation.

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

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