PSYCHOMETRIKA—VOL. 87, NO. 2, 432–476
JUNE 2022
https://doi.org/10.1007/s11336-021-09803-z

A SYSTEMATIC STUDY INTO THE FACTORS THAT AFFECT THE PREDICTIVE ACCURACY OF MULTILEVEL VAR(1) MODELS

GINETTE LAFIT, KRISTOF MEERS AND EVA CEULEMANS

KU LEUVEN – UNIVERSITY OF LEUVEN

The use of multilevel VAR(1) models to unravel within-individual process dynamics is gaining momentum in psychological research. These models accommodate the structure of intensive longitudinal datasets in which repeated measurements are nested within individuals. They estimate within-individual auto- and cross-regressive relationships while incorporating and using information about the distributions of these effects across individuals. An important quality feature of the obtained estimates pertains to how well they generalize to unseen data. Bulteel and colleagues (Psychol Methods 23(4):740–756, 2018a) showed that this feature can be assessed through a cross-validation approach, yielding a predictive accuracy measure. In this article, we follow up on their results, by performing three simulation studies that allow to systematically study five factors that likely affect the predictive accuracy of multilevel VAR(1) models: (i) the number of measurement occasions per person, (ii) the number of persons, (iii) the number of variables, (iv) the contemporaneous collinearity between the variables, and (v) the distributional shape of the individual differences in the VAR(1) parameters (i.e., normal versus multimodal distributions). Simulation results show that pooling information across individuals and using multilevel techniques prevent overfitting. Also, we show that when variables are expected to show strong contemporaneous correlations, performing multilevel VAR(1) in a reduced variable space can be useful. Furthermore, results reveal that multilevel VAR(1) models with random effects have a better predictive performance than person-specific VAR(1) models when the sample includes groups of individuals that share similar dynamics.

Key words: intensive longitudinal data, linear mixed effect models, cross-validation, multicollinearity, principal components.

1. Introduction

The popularity of data collection approaches such as the experience sampling method (Larson and Csikszentmihalyi 1983; Myin-Germeys et al. 2018) and Ambulatory Assessment (Trull and Ebner-Priemer, 2013) has increased the availability of intensive longitudinal datasets. These datasets allow studying the dynamics of psychological functioning within individuals over time (Molenaar, 2004), as is increasingly done in many research areas. For example, in psychopathology research, numerous researchers now focus on the vicious dynamic interactions between individual symptoms (e.g., Borsboom and Cramer 2013; Bringmann et al. 2016). In affect research, within-person analyses are used to study emotional inertia, sensitivity, augmenting and blunting (e.g., Krone et al. 2018; Kuppens et al. 2010; Kuppens et al. 2012; Pe et al. 2015).

A widely used statistical approach to model these within-person dynamics is the lag-one vector autoregressive (VAR(1)) model (see, e.g., Bringmann et al. 2016; Hamaker et al. 2015; Pe et al. 2015; Wichers 2014; Wigman et al., 2015), comprising the lag-one autoregressive (AR(1)) model as a special case (see, e.g., Jongerling et al. 2015; Kuppens et al. 2010; Liu 2017). In this model, which was first proposed in the econometric literature (see, e.g., Lütkepohl 2005), each variable is regressed upon all variables (including itself) at the previous time point. Therefore, for each particular variable, the model allows estimating the effect of its past values on current

Correspondence should be made to Ginette Lafit, Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium. Email: ginette.lafit@kuleuven.be

values (i.e., autoregressive effects) as well as the effect of past values of the rest of the variables (i.e., cross-regressive effects). This model can be estimated at the individual level (i.e., person-specific VAR(1)) or using a multilevel (see, e.g., Asparouhov et al. 2018; Bulteel et al. 2018a; Bringmann et al. 2013; Bringmann et al. 2016; Jongerling et al. 2015) or clustering (e.g., Bulteel et al. 2016a; Ernst et al. 2019) framework. Within the multilevel approach, the autoregressive and cross-regressive effects can be considered as random variables that vary across individuals. This allows to estimate the within-individual autoregressive and cross-regressive effects while incorporating and using information about the distribution of these effects across individuals. The clustering approaches try to classify the individuals in mutually exclusive groups, based on their auto- and cross-regressive effects.

Although very popular and useful at first sight, the application of the VAR(1) model is challenging, for three reasons. Firstly, in VAR(1) models there are a large number of parameters to be estimated. Specifically, the number of parameters amounts to P^2, where P denotes the number of variables. Given that psychological intensive longitudinal datasets usually include 20 to 120 participants with 20 to 200 repeated measurements (e.g., Morren et al. 2009; Ono et al. 2019; Vachon et al. 2019), and the number of variables included in the VAR(1) model generally includes between 2 and 12 variables (e.g., Bringmann et al. 2013; Bringmann et al. 2016; Mansueto et al. 2020), the question rises whether such data contain enough information to reliably estimate so many parameters, and more generally, how estimation performance is affected by the number of measurements, the number of persons, and the number of variables.

Secondly, in psychological applications, the variables in intensive longitudinal datasets can be highly contemporaneously correlated (e.g., Pe et al. 2015; Sels et al. 2016). The presence of contemporaneous correlations is obviously implied by the auto- and cross-regressive effects of the variables and the correlations between the innovations. However, in some cases, part of the observed contemporaneous correlation might simply be due to the design of the momentary questionnaires. Indeed, momentary assessments often include some items that show clear content overlap. For instance, negative affect can be conceptualized as a latent construct that can be captured well using several indicator variables (e.g., Brose et al. 2015; Merz and Roesch 2011; Zautra et al. 2005). These indicators are sometimes almost synonyms (e.g., sad and down) and may even be considered redundant in specific cases (e.g., irritated and irritable) (see Eisele et al. 2020). Such content overlap thus adds to the contemporaneous correlations that may be expected based on the hypothesized VAR(1) dynamics, and may lead to highly unstable VAR(1) coefficients, as demonstrated by Bulteel et al. (2018b). Moreover, the shared dynamics (i.e., the common effects of two or more variables on future states of an outcome on top of the current state of the outcome and the effects of other variables; see, e.g., Bulteel et al. 2016b) between the variables are not directly reflected in the model. A promising solution to overcome these problems consists of applying an exploratory dimension reduction approach. For example, Bulteel et al. (2018b) proposed to fit VAR(1) models in a reduced variable space. In this approach the variables are first reduced to a few components, by means of principal component analysis. Next, the VAR(1) model is estimated on these components rather than on the original variables. However, it is still unclear how the performance of such a dimension reduction-based procedure is affected by the number of extracted components, as well as by the number of variables, persons, and measurements. Moreover, so far, this approach has not been extended to the multilevel setting.

Thirdly, many research questions focus on individual differences in the within-person dynamics. Existing VAR(1) approaches allow to capture such differences in different ways. As we indicated above, one can model the data of each person separately or use a multilevel approach. Moreover, a number of clustering techniques have been proposed as well (e.g., Bulteel et al. 2016a; Ernst et al. 2019), that can be useful in case one assumes the individual differences to be at least partly categorical. Here, the question becomes how to know which approach captures the individual differences well and how this is affected by the previously mentioned data characteristics? Moreover, can we also disclose correctly that sometimes individual differences are negligible, calling for fixed rather than random effects?

Up to now, relatively few studies (see, e.g., Krone et al. 2016; Krone et al. 2017; Liu 2017; Mansueto et al. 2020; Schultzberg and Muthén 2018) have investigated how the performance of VAR(1) models is affected by different data characteristics. Moreover, these studies employed different performance measures. In the general statistical literature, two popular performance criteria are estimation accuracy and predictive accuracy. Estimation accuracy assesses whether the statistical approach can retrieve the parameter values of the true underlying model (i.e., absence of bias) and quantifies how much parameter estimates vary across different samples from the same population model (i.e., standard error) (see, e.g., Friedman et al. 2001). Evaluating estimation accuracy when fitting the true model or a misspecified one makes most sense when the true parameters of the model are known. Moreover, estimation accuracy is then directly related to the quality of statistical inferences regarding these parameters. In contrast, predictive accuracy allows investigating how well one or more obtained statistical models for a particular dataset generalize as a whole (i.e., the combination of all parameter values) to unseen data, without even claiming that the true model is among the considered ones. When focusing on the true model only, better estimation accuracy (due to among others larger sample size) is expected to lead to better predictive accuracy. However, when also including incorrect models, there are cases in which using the correct (rather than an incorrect) model can lead to worse predictive accuracy (Yarkoni and Westfall, 2017). For example, when the correct model includes a large number of parameters and sample size is small, this usually leads to overfitting problems, implying that the estimated parameters do not predict new unseen data from the same individuals well (see, e.g., Babyak 2004). Therefore, applying a more simple, but incorrect model, may lead to less overfitting, and therefore to better predictive accuracy, whereas estimation accuracy is obviously low.

Published studies on the performance of VAR(1) models have primarily focused on the estimation accuracy of person-specific and multilevel autoregressive (AR) models, rather than on VAR(1) (e.g., Krone et al. 2016, Krone et al. 2017; Liu 2017). AR models are restricted and thus more simple versions of VAR since they estimate the dynamics of a variable by regressing current values on past values of the same variable only, ignoring past values on other variables. These studies show that the Bayesian and maximum likelihood-based estimation procedures yield the best performance and that estimation accuracy increases with the number of individuals and the number of measurement occasions. Liu (2017) investigated the performance of the estimation accuracy of AR estimates and one-step ahead predictions and showed that multilevel approaches outperform person-specific ones, unless the models are incorrectly specified (e.g., incorrect lag orders). Hence, it remains largely unclear how the estimation performance of the more general and complex VAR(1) approaches depends on the number of measurements, persons, and variables, on the correlations between the variables, and on whether and how the effects differ across individuals.

Although estimation accuracy is a useful quality feature, it does not readily reveal whether (some of) the VAR(1) models considered can generalize well to unseen data given the typical intensive longitudinal datasets in psychology. To get more insight into this quality aspect, Bulteel et al. (2018a) were the first to conduct a study on the predictive accuracy of the VAR(1) model. These authors investigated the predictive performance of person-specific AR(1) and VAR(1) models, multilevel AR(1) and VAR(1) models and person-specific lasso VAR(1) models. This was done via blocked cross-validation techniques that estimate the out-of-sample mean squared prediction error of the different models. Blocked cross-validation allows taking the serial dependency within the time series into account, by dividing the dataset into blocks of consecutive measurements. By reanalyzing three psychological datasets, Bulteel et al. (2018a) show that the predictive accuracy of person-specific VAR(1) model was lower than that of multilevel AR(1) and VAR(1) models. This result raises the question if person-specific VAR(1) models can adequately capture the regularities of longitudinal data typically collected in the field of psychology, allowing to predict future states. Furthermore, they show that the multilevel AR(1) and VAR(1) models have, in general, the best overall predictive performance. These results indicate that pooling information across individuals using a multilevel framework prevents overfitting to at least some extent and improves predictive accuracy. Since Bulteel et al. (2018a) studied predictive accuracy using empirical datasets, they had less grip on the underlying data characteristics. However, as introduced above, we are especially interested in the influence of five data characteristics. First, the number of measurements per person defines the amount of available information, with higher numbers usually leading to less overfitting, given a model with a given number of parameters. Second, higher numbers of individuals may allow for better modeling of individual differences. Third, the number of variables directly impacts the number of parameters to be estimated and can be expected to lower predictive accuracy. Fourth, the collinearity between the variables decreases the distinguishability of the effects and thus may increase overfitting. Finally, the distributional shape of the individual differences in the VAR(1) parameters (e.g., normal distribution versus multimodal distributions) might be important as well, in that model misspecifications may lower the predictive accuracy.

Based on this short review, we conclude that current literature has not provided full answers to the three challenges and associated questions introduced above and that complementing the estimation accuracy perspective with a predictive accuracy perspective is worthwhile. Therefore, we implement a simulation-based approach to systematically investigate the effect of these five data characteristics on the predictive accuracy of person-specific and multilevel VAR(1) models. The simulation approach uses a number of population models and concrete specifications of the associated parameters and data characteristics to generate a large number of datasets. Each of these datasets is then analyzed using different model specifications. Because the true model is known, the simulation-based procedure sheds light on how data characteristics influence predictive and estimation accuracy, and can be used to compare the performance of correct and incorrect model specifications.

Specifically, in the present article, we conduct three simulation studies. In the first study, we investigate the predictive accuracy of different model specifications when we manipulate the number of persons, the number of time points, and the number of variables. We consider six model specifications. The first two specifications define person-specific AR(1) and VAR(1) models. The third and fourth pertain to a multilevel AR(1) model, with the autoregressive effect being fixed across participants in the first specification and random in the second. The fifth and sixth specification comprises a multilevel VAR(1) model, with the difference between the two specifications being again the fixed versus random nature of the autoregressive and cross-lagged effects. In the second study, we investigate the effect of strong contemporaneous correlations between the variables on the predictive performance of person-specific and multilevel VAR(1) models. Moreover, we evaluate the performance of an estimation procedure that handles multicollinearity by first reducing the number of variables through a dimension reduction approach (Bulteel et al. 2018b). In the third study, we focus on the predictive accuracy of person-specific and multilevel VAR(1) models when individual differences have a multimodal distribution.

The remainder of the article is organized as follows. First, we introduce the person-specific and multilevel VAR(1) model approach. Also, we present a modeling procedure based on dimensionality reduction to handle multicollinearity in multilevel VAR(1) models. Next, we detail performance measures that can be used to assess predictive accuracy. In the following sections, we study the predictive performance of person-specific and multilevel VAR(1) models under different data characteristics by conducting three simulation studies. We conclude with a summary of the findings and discuss future research extensions.

2. Vector Autoregressive Models

In this section, we present the models under study. We will apply these models to intensive longitudinal datasets, that each include N persons. We assume that, for each person i, P variables were observed at T_i equidistant measurement occasions.

2.1. *The Person-Specific VAR(1) and AR(1) Models*

In the person-specific VAR(1) model, the data of each person are modeled separately. Each variable is regressed on all variables (including itself) at the previous measurement occasion. Hence, for person i ($i = 1, \ldots, N$), the ($P \times 1$) vector $\mathbf{Y}_{it}$, where $t = 1, \ldots, T_i$, is modeled as follows

$$\mathbf{Y}_{it} = \mathbf{c}_i + \mathbf{\Psi}_i \mathbf{Y}_{it-1} + \boldsymbol{\epsilon}_{it} \tag{1}$$

where $\mathbf{Y}_{it-1}$ is the vector of variable scores at the previous measurement occasion for person i. $\mathbf{c}_i$ is a ($P \times 1$) vector holding the person-specific intercepts. The ($P \times P$) matrix $\mathbf{\Psi}_i$ is the person-specific transition matrix. This matrix holds the regression coefficients, which quantify how current states depend on previous states through autoregressive (i.e., diagonal elements) and cross-regressive (i.e., off-diagonal elements) unique direct effects. Shared effects of the previous states are not reflected in the regression coefficients (Bulteel et al. 2016b). P AR(1) models are obtained when all cross-regressive effects are set to zero, implying that their combined transition matrix is a diagonal matrix and that each variable is only regressed on itself. The errors $\boldsymbol{\epsilon}_{it} = (\epsilon_{it1}, \ldots, \epsilon_{itP})^T$, which are also often called innovations, represent unmodeled influences that occur on a timespan shorter than the lag, that is, the part of the data that cannot be predicted by past observations. We assume that these errors are multivariate Gaussian distributed with mean zero and covariance matrix $\mathbf{\Sigma}_{\epsilon}$.

Furthermore, we impose that for a person i the process $\mathbf{Y}_i$ is stationary, which means that the mean, variance, and autocorrelation structure is time-invariant (Hamilton, 1994). For person i, one can investigate whether the associated VAR(1) model is stationary by verifying that the modulus of the eigenvalues of the transition matrix $\mathbf{\Psi}_i$ is smaller than one (Lütkepohl 2005). For the AR(1) model, this condition implies that the absolute value of each AR coefficient individually has to be smaller than 1.

The person-specific VAR(1) model is often estimated by fitting separate ordinary least square regressions for each variable, or by means of maximum likelihood estimation (see, Hamilton 1994; Lütkepohl 2005). Previous research showed that both estimation procedures have identical asymptotic properties (Lütkepohl, 2005). The dynamic structural equation modeling approach implemented in MPlus (Muthén and Muthén, 2009), in which the full model is estimated at once, is rapidly gaining popularity as well (Asparouhov et al., 2018). For three reasons listed in the next subsection, we adopted a single equation framework in this paper, and used the ordinary least square regression approach to estimate person-specific VAR(1) models.

2.2. *The Multilevel VAR(1) and AR(1) Models*

The intensive longitudinal datasets that we consider in this paper have a multilevel structure in that the measurement occasions are nested within individuals. Multilevel or linear mixed-effect regression models (Goldstein, 2011) were developed to effectively handle multilevel data. It therefore makes sense that multilevel extensions of the VAR(1) model have been proposed to analyze intensive longitudinal data, using either a single-equation (Bringmann et al., 2013) or dynamic structural equation modeling framework (see Asparouhov et al. 2018; Jongerling et al. 2015; McNeish and Hamaker 2020). These multilevel VAR(1) models allow capturing individual differences by including random autoregressive and cross-regressive effects that vary across persons. Yet, each of the autoregressive and cross-lagged effects can also be fixed across persons, implying that they have the same value across individuals.

In a random effects multilevel VAR(1) model, the vector of intercepts $\mathbf{c}_i$ and the transition matrix $\boldsymbol{\Psi}_i$ in Eq. (1) are considered random variables:

$$\begin{aligned}\mathbf{c}_i &= \mathbf{c}^g + \boldsymbol{\nu}_i\\ \boldsymbol{\Psi}_i &= \boldsymbol{\Psi}^g + \boldsymbol{\Upsilon}_i\end{aligned}$$

where $\mathbf{c}^g$ is the $(P \times 1)$ vector of fixed intercepts and $\boldsymbol{\nu}_i$ represents the random deviations of the person-specific intercepts of person i from these fixed intercepts. $\boldsymbol{\Psi}^g$ is the fixed $(P \times P)$ transition matrix, and $\boldsymbol{\Upsilon}_i$ is a $(P \times P)$ matrix of person-specific random deviations from the fixed transition matrix. The random effects represented by the $(P(1+P) \times 1)$ vector $(\text{vec}(\boldsymbol{\nu}_i), \text{vec}(\boldsymbol{\Upsilon}_i)^T$, are usually assumed to be multivariate Gaussian distributed with a mean of zero and a $(P(1+P) \times P(1+P))$ covariance matrix $\boldsymbol{\Sigma}_\nu$. The model assumes that the random effects are uncorrelated with the within-individual errors $\boldsymbol{\epsilon}_{it}$. These within-individual innovations $\boldsymbol{\epsilon}_{it}$ are assumed to be Gaussian distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_\epsilon$. Although we will only consider fully fixed or fully random settings, researchers can in principle fit alternative models in which some regression effects are fixed and others are random, however raising the question how one should decide which effects can be fixed (see, e.g., Bar et al. 2013; Bates et al. 2015a; Clark and Linzer 2015; Gelman 2005). P multilevel AR(1) models are obtained when the transition matrix $\boldsymbol{\Psi}_i$ is restricted to a diagonal matrix.

Compared to person-specific modeling, multilevel modeling shrinks the person-specific regression weights toward the population average or fixed effects, since extreme values occur only rarely under the Gaussian assumption (Hox, 2010). The shrinkage biases the estimates of the person-specific effects but decreases their variance, and therefore better controls for overfitting than person-specific VAR(1) models (e.g., Bulteel et al. 2018a), leading to better predictive accuracy. The amount of shrinkage will depend on the number of measurements per person, with less measurements per person leading to more shrinkage.

In this paper, in line with Bringmann et al. (2013), Bulteel et al. (2018a) and Liu (2017), we estimate multilevel VAR(1) models by separately fitting a linear-mixed effect regression for each variable in the model (i.e., single equation approach). We use a restricted maximum likelihood procedure, which yields unbiased estimates of the variance components (see, e.g., Raudenbush and Bryk 2002). Specifically, to estimate linear mixed-effect models we use the open-source software Julia (Bezanson et al. 2017) package MixedModels.jl (Bates et al. 2016). This package provides a similar functionality to the R (R Core Team 2020) package lme4 (Bates et al. 2015b), but is computationally more efficient. We resort to this approach rather than to dynamic structural equation modeling because of three reasons. First, the single-equation framework allows estimating models that include a larger number of predictors and maximal random effect structures. Indeed, when trying out dynamic structural equation modeling on some of our simulated datasets, we ran into estimation problems for larger numbers of variables. Second, the current dynamic structural equation framework does not provide predictive accuracy measures, which is the main focus of this paper. Third, we aim to provide freely available code that can be used in open-source software to reproduce and extend the analyses presented in this paper. So far, dynamic structural equation modeling is implemented exclusively in Mplus, a closed-source and licensed software. Because of the second reason, the comparison between the predictive accuracy of the single-equation framework and dynamic structural equation modeling remains an empirical question.

2.3. Dimension Reduction-Based VAR(1) and AR(1) Models

As discussed in the introduction of this paper, it often happens in intensive longitudinal research that the variables under study are highly contemporaneously correlated. In regression analysis and thus also in VAR(1), such multicollinearity makes it difficult to distinguish the unique contribution of a predictor to the variance of the outcome variable and results in highly unstable parameter estimates (Cohen et al. 2013). To overcome the issue of multicollinearity in the estimation of person-specific VAR(1) models, Bulteel et al. (2018b) proposed a two-step procedure, called PC-VAR(1). In the first step, the variables are reduced to a few principal components by means of principal component analysis (PCA) and these principal components are rotated toward a simple structure. In the second step, the person-specific VAR(1) model is estimated on these rotated components rather than on the original variables. In terms of predictive accuracy, Bulteel et al. 2018b) showed that when variables are highly contemporaneously correlated, PC-VAR(1) outperforms the person-specific VAR(1) model, as well as Lasso-based approaches.

To analyze the data of multiple individuals simultaneously, the PC-VAR(1) approach can be extended using simultaneous component analysis (SCA) (Timmerman and Kiers, 2003) in the first step and fitting a person-specific (V)AR(1) model or a multilevel (V)AR(1) model to the resulting component scores. We use SCA because it models the intra-individual contemporaneous associations between the variables only. To implement this approach, first, all lagged variables (i.e., the predictors) are person-mean centered. This centering step discards interindividual differences in means and thus allows to focus on the within-person contemporaneous correlations (a further explanation can be found in Ceulemans et al. 2016), when performing dimension reduction. Second, the person-mean centered predictors are scaled using the standard deviation of the whole sample (i.e., grand standard deviation). This scaling step makes sure that every predictor gets the same weight in the dimension reduction step.

We used the least restrictive SCA approach, often referred to as SCA-P (e.g., Kiers and ten Berge 1994a; Timmerman and Kiers 2003). SCA-P boils down to implementing PCA on the person-centered and scaled data of all participants simultaneously:

$$\mathbf{Y}_{it-1} = \mathbf{\Gamma}\mathbf{F}_{it-1} + \boldsymbol{\varepsilon}_{it-1} \tag{2}$$

where $\mathbf{\Gamma}$ denotes a $(P \times Q)$ loading matrix, which is assumed to be the same across the persons. $\mathbf{F}_{it-1}$ is a $(Q \times 1)$ vector that holds the component scores of person i at time point $t-1$. $\boldsymbol{\varepsilon}_{it-1}$ is a $(P \times 1)$ vector of measurement errors of person i at time $t-1$. Note that unlike innovations, the effect of measurement errors is not carried over to the next time point (see Schuurman and Hamaker 2019).

An important decision pertains to the number of components Q. This decision can be based on theory or interpretation. Alternatively, automated data-driven procedures can be applied, such as the CHull procedure (Ceulemans and Kiers 2006; Wilderjans et al. 2013), parallel analysis (e.g., Crawford et al. 2010; Horn 1965), or cross-validation (Bulteel et al. 2018b). Given its computational speed and its excellent performance in a diverse set of simulation studies (see Ceulemans and Kiers 2009; Ceulemans et al. 2011; Ceulemans and Van Mechelen 2005; Lorenzo-Seva et al. 2011; Schepers et al. 2008), we opted to use the CHull procedure in this paper. The CHull procure automatizes the well-known scree test of Cattell (1966), by computing scree test values

$$st_Q = \frac{VAF_Q - VAF_{Q-1}}{VAF_{Q+1} - VAF_Q} \tag{3}$$

for SCA solutions with different numbers of components Q with $Q = 1 \ldots P$ and picking the Q-value that leads to a maximal st_Q-value. Note that VAF denotes the percentage of variance explained. Setting VAF_0 to zero, all solutions can be selected, except for the one where $Q = P$, which makes sense as this solution would not imply any reduction at all.

The extracted Q principal component scores are uncorrelated and restricted to have a variance of one, implying that the loadings express the correlations between the P predictors and the Q components and thus allow to label the components. However, labeling the principal components is often challenging because all variables have non-negligible loadings. Therefore, in line with Bulteel et al. (2018b), we propose to rotate the components using Varimax rotation (see Kiers and ten Berge (1994b). This orthogonal rotation procedure aims for simple structure, implying there is one high loading for each variable and the rest of them are close to zero.

Alternative dimension reduction approaches have been proposed to estimate multilevel VAR(1) models on data in which the variables are highly contemporaneously correlated, in a simultaneous rather than two-step fashion. These approaches usually pertain to dynamic factor analysis techniques. For example, Song and Zhang (2014) proposed a confirmatory multilevel dynamic factor model to model lagged relationships between latent constructs at the within-person level. These simultaneous approaches take a confirmatory stance, starting from prior hypotheses about which variables should load on which latent constructs and restricting the fitted model accordingly. In contrast, our SCA-based approach is an exploratory one and can thus be used when no clear prior hypotheses are available. Moreover, the SCA-P step has a closed form solution based on singular value decomposition, making it computationally inexpensive. Finally, note that dynamic factor models are specific instances of dynamic structural equation models. We have explained in the previous section why we did not consider this modeling framework in the present paper.

3. Predictive Accuracy

3.1. Performance Measures to Evaluate Predictive Accuracy

A widely used method to assess the predictive accuracy of estimated models for unseen data is cross-validation (CV) (e.g., Bulteel et al. 2018b). In a CV framework, the observed data are split up into a training and test set. The models of interest are first fitted to the training set, and the estimated parameters are then used to predict the observations in the test set. Finally, the predictive accuracy of the model is estimated by using the predicted values for the test sets to compute a measure such as the mean squared error of the prediction errors (MSPE) (Friedman et al. 2001).[1]

In the context of intensive longitudinal data, Bulteel et al. (2018a) investigated the performance of different CV procedures (i.e., leave-one-out CV, K-fold CV, K-blocked CV, and

[1] Next to the mean squared prediction error (MSPE), we also used the out-of-sample coefficient of determination R^2 (Campbell and Thompson, 2008) to assess the predictive accuracy of the considered models. For a variable j, the out-of-sample R_j^2 represents the proportion of variance of Y_j that is explained by the predictive model, estimated using other data. In our cross-validation context, the out-of-sample R_j^2 can be computed as follows:

$$\mathrm{R}_j^2 = 1 - \frac{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T_i^{(k)}}(Y_{itjk} - \hat{Y}_{itjk})^2}{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T_i^{(k)}}(Y_{itjk} - \bar{Y}_{itjk})^2} \tag{4}$$

where $\hat{Y}_{itjk}$ indicates the predicted values based on the training data, and $\bar{Y}_{itjk}$ is the mean of variable j. The overall R^2 is computed as the average of the P R_j^2-values. The out-of-sample R^2 takes values in the interval $[-\infty, 1]$. Negative values reflect that the predictive model has a poor predictive performance, whereas a value equal to one shows that the model perfectly predicts unseen data. Obviously, in most cases, we expect R^2 to be smaller than the in-sample counterpart. Since this measure shows the same patterns as the MSPE, we did not include the results in the paper, but interested readers can consult them in the supplementary material: https://osf.io/rs6un/.

hv-blocked CV) for assessing the predictive accuracy of person-specific and multilevel VAR(1) and AR(1) models. The procedures differed in how the training and test sets are constructed. The authors showed that K-blocked CV outperformed the other procedures. This makes sense as K-blocked cross-validation takes the serial dependency in time series into account, by dividing the data of each person into K blocks of consecutive measurements. Each block is consecutively selected as the test set, and the remaining $K-1$ blocks are used as the training set.

In this paper we follow this K-blocked CV approach. Whereas this is rather straightforward for the person-specific and multilevel models (see Bulteel et al. 2018a), some additional choices had to be made for the dimension reduction-based approaches. Here, we apply all steps discussed in Sect. 2—preprocessing, extraction of components, rotation on the training data, using the true number of components. Except for the case where we extract a single component only, we always rotate the components to simple structure using a Varimax rotation. Next, we turn to the test set, and preprocess the variables in the test set using the person means and grand standard deviation of the training set. We compute the rotated component scores of the test set using the transposed pseudo-inverse of the rotated loading matrix from the training set. These rotated component scores are then multiplied by the transition matrix and the rotated loadings of the training set, yielding the predicted test set values.

Once we computed the predicted values for the test set, we can calculate the MSPE. To this end, the differences between the observed and predicted values in the test set are squared and averaged over test sets, for each variable Y_j $(j = 1, \ldots, P)$:

$$\text{MSPE}_j = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i^{(k)}}\sum_{t=1}^{T_i^{(k)}}\left(Y_{itjk} - \hat{Y}_{itjk}\right)^2\right) \tag{5}$$

where $T_i^{(k)}$ is the total number of measurement occasions of a person i in block k. Y_{itjk} indicates the t-th observation of variable j for person i in block k. $\hat{Y}_{itjk}$ is the predicted value based on the parameter estimates for the training set. Finally, the overall MSPE is computed as the average of the P MSPE_j-values:

$$\text{MSPE} = \frac{1}{P}\sum_{j=1}^{P}\text{MSPE}_j. \tag{6}$$

3.2. *Relating Predictive Accuracy to Estimation Accuracy*

As mentioned in the introduction, the relation between estimation accuracy and predictive accuracy is not that straightforward. To better demonstrate this point, we assume the test set data $\mathbf{Y}_{it}^{(k)} = f(\mathbf{Y}_{it-1}^{(k)}) + \boldsymbol{\epsilon}_{it}$, where $f(\mathbf{Y}_{it-1}^{(k)}) = \mathbf{c}_i + \boldsymbol{\Psi}_i\mathbf{Y}_{it-1}^{(k)}$ is the target function, specified by the true model underlying the test set. This target function is then estimated based on the training data $\mathbf{Y}_{it}^{(-k)}$ for that test set, consisting of all CV blocks except k. Predicting the test set through the estimated target function yields $\hat{f}(\mathbf{Y}_{it-1}^{(k)})$. The expected MSPE given the estimated target function can then be written as follows (see Friedman 1997):

$$E\left[\mathbf{Y}_{it}^{(k)} - \hat{f}\left(\mathbf{Y}_{it-1}^{(k)}\right)\right]^2 = E\left[f\left(\mathbf{Y}_{it-1}^{(k)}\right) - \hat{f}\left(\mathbf{Y}_{it-1}^{(k)}\right)\right]^2 + E\left[\boldsymbol{\epsilon}_{it}\right]. \tag{7}$$

The second term reflects the innovation variance within the test set, which is irreducible because it does not depend on how well the estimated target function approximates the true target function. The first term, however, which focuses on the mean squared difference between the prediction based on the true target function and that based on the estimated one, can be further decomposed as follows:

$$
\begin{aligned}
E\left[f\left(\mathbf{Y}_{it-1}^{(k)}\right)-\hat{f}\left(\mathbf{Y}_{it-1}^{(k)}\right)\right]^{2}=&\left[f\left(\mathbf{Y}_{it-1}^{(k)}\right)-E\left[\hat{f}\left(\mathbf{Y}_{it-1}^{(k)}\right)\right]\right]^{2}\\
&+E\left[\hat{f}\left(\mathbf{Y}_{it-1}^{(k)}\right)-E\left[\hat{f}\left(\mathbf{Y}_{it-1}^{(k)}\right)\right]\right]^{2},
\end{aligned}
\tag{8}
$$

effectively splitting it into a mean squared bias of the predictions based on the estimated target function, and their variance, both considered across different possible sets of training data. The latter is hypothetical obviously, because we have only one set of training data in our CV setting. When the estimated target function is very sensitive to the values included in the training set, which can be expected when complex target functions are combined with small sample sizes, the variance of the predicted values based on different training sets is expected to be large.

This variance and bias decomposition of the MSPE sheds some light on how estimation and predictive accuracy are related: to obtain good predictive accuracy, the estimated target functions should be characterized by small bias as well as small variance. This is why estimating a true but complex target function may in the end yield worse predictive accuracy than estimating an incorrect but simple target function. Indeed, the first situation may lead to a lot of variance despite small bias, whereas the second may imply small variance on top of a moderate bias, making the second target function, though incorrect, the winner. We may, for instance, observe this result, when the correct target function is a person specific VAR(1), and the incorrect ones are a person-specific AR(1) or a multilevel VAR(1). Here, we expect that the incorrect models introduce estimation bias while their predictive accuracy outperforms that of the correct one because the lower number of parameters or the shrinkage of the person-specific effects toward the population mean imply lower variance.

4. Study I: The Effect of the Number of Measurement Occasions, the Number of Persons, and the Number of Variables

The goal of the first simulation study is to investigate the effect of three data characteristics on the predictive accuracy of person-specific and multilevel AR(1) and person-specific and multilevel VAR(1) models. Specifically, we focus on the effect of the number of measurement occasions, the number of persons, and, in case of the VAR(1) models, the number of variables. We predict that the number of measurement occasions and the number of persons are positively related to predictive accuracy, while the number of variables has a negative impact. The datasets are generated according to six models: person-specific AR(1) models (AR), person-specific VAR(1) model (VAR), multilevel AR(1) model with fixed autoregressive effects (MAR.FE), multilevel AR(1) model with random autoregressive effects (MAR.RE), multilevel VAR(1) model with a fixed transition matrix (MVAR.FE), and multilevel VAR(1) model with random transition matrices (MVAR.RE). Throughout the simulation study, we assume that the person-specific intercepts are equal to zero, allowing us to fully assess the impact of the regressive effects and how they are handled by the different models.

4.1. Simulation Design

The datasets are generated according to a four-factorial design, including 100 replicates per design cell. The first three factors pertain to the three above-mentioned data characteristics. Specifically, we manipulate: (i) the number of individuals: 20, 60, and 120; (ii) the number of measurement occasions per individual: 50, 100, and 200; and (iii) the number of variables: 2, 4, and 8. The fourth factor pertains to the data generating model: AR, VAR, MAR.FE, MAR.RE, MVAR.FE, and MVAR.RE. With respect to the third factor—number of variables, because of the large number of analyses that have to be run, we only vary the number of variables when generating data with VAR, MVAR.FE, and MVAR.RE. For the cases in which the underlying model is AR, MAR.FE, and MAR.RE, we always set the number of variables to four. This decision is motivated by the fact that the complexity of AR(1) models increases linearly with the number of variables, whereas for VAR(1) models the number of transition parameters equals the squared number of variables. This means that we generated 108 conditions, with 100 replicates per condition, yielding 10800 datasets.

We will now elaborate on how we generated the data for the six models under consideration. To briefly investigate estimation accuracy (see "Appendix") also, we did not vary the regressive parameters within each design cell.[2] The value of the auto- and cross-regressive effects are based on previous simulation studies (see Bulteel et al. 2016a; Ernst et al. 2019; Liu 2017) and empirical applications of these models to intensive longitudinal data (see Krone et al. 2018; Kuppens et al. 2010). Throughout the simulation settings, we set the standard deviation of the within-individual innovations σ_ϵ to one for all individuals and the covariances to 0.2. Moreover, for the person-specific and multilevel VAR(1) models, to guarantee that the N person-specific and the fixed effects transition matrices are stationary, the person-specific and fixed regression weights were multiplied by $0.99/(|\lambda_p|)$, where $|\lambda_p|$ denotes the absolute value of the maximum eigenvalue of the person-specific transition matrix.

When simulating from a person-specific AR(1) model, we generate N autoregressive effects for each person separately from a uniform distribution on the interval [0.2, 0.6]. To obtain person-specific VAR(1) data, we sample the N person-specific autoregressive effects from a uniform distribution on the interval [0.2, 0.6] and the person-specific cross-regressive effects from a uniform distribution on the interval [0.05, 0.2].

When the data generating process is a MAR.FE model, half of the fixed autoregressive effects are set to 0.3, and the other half to 0.4. In case we simulate data with a MAR.RE model, we set the fixed autoregressive effects as in the MAR.FE setting. The random effects are drawn from a multivariate Gaussian distribution, with all means and covariances set to zero and all variances fixed to 0.1 (see, e.g., Liu 2017). To ensure that the generated time series of each variable and each person conformed to the stationary assumption, we checked whether the person-specific autoregressive effects had an absolute value smaller than one. Therefore, the distribution of each random autoregressive effect approximates a truncated Gaussian distribution (Ernst et al. 2019).

In the MVAR.FE setting, half of the P fixed autoregressive effects are set to 0.3, and the other half to 0.4. The $P(P-1)$ fixed cross-regressive effects are sampled from a uniform distribution on the interval [0.05, 0.2]. We again guaranteed that the generated time series are stationary. The resulting transition matrices are given by:

Transition matrix when $P = 2$

$$\boldsymbol{\Psi}^g = \begin{bmatrix} 0.300 & 0.111 \\ 0.168 & 0.400 \end{bmatrix} \tag{9}$$

[2] In a previous version of the manuscript, we conducted the analysis varying the regressive parameters within each design cell. The simulation results are included in the OSF page of the project and are comparable with the ones presented in this paper.

Transition matrix when $P = 4$

$$\boldsymbol{\Psi}^g = \begin{bmatrix} 0.400 & 0.133 & 0.118 & 0.065 \\ 0.057 & 0.300 & 0.152 & 0.185 \\ 0.129 & 0.118 & 0.300 & 0.087 \\ 0.184 & 0.194 & 0.136 & 0.400 \end{bmatrix} \tag{10}$$

Transition matrix when $P = 8$

$$\boldsymbol{\Psi}^g = \begin{bmatrix} 0.320 & 0.086 & 0.093 & 0.086 & 0.135 & 0.056 & 0.051 & 0.098 \\ 0.065 & 0.320 & 0.131 & 0.114 & 0.052 & 0.118 & 0.096 & 0.147 \\ 0.055 & 0.073 & 0.240 & 0.082 & 0.092 & 0.081 & 0.101 & 0.150 \\ 0.130 & 0.138 & 0.116 & 0.240 & 0.158 & 0.119 & 0.112 & 0.113 \\ 0.147 & 0.094 & 0.125 & 0.053 & 0.240 & 0.078 & 0.080 & 0.089 \\ 0.085 & 0.137 & 0.040 & 0.069 & 0.147 & 0.320 & 0.099 & 0.058 \\ 0.120 & 0.138 & 0.097 & 0.120 & 0.146 & 0.063 & 0.240 & 0.152 \\ 0.051 & 0.135 & 0.066 & 0.090 & 0.061 & 0.134 & 0.155 & 0.320 \end{bmatrix} \tag{11}$$

In the MVAR.RE scenario, the fixed autoregressive and cross-regressive effects are determined as in the MVAR.FE scenario. The random effects are generated from a multivariate Gaussian distribution, with all means set to zero, and with a diagonal covariance matrix in which the variances are set to 0.025 (see, e.g., Ernst et al. 2019). We again ensure that the generated time series of each participant conformed to the stationary assumption by checking the eigenvalues of the individual transition matrices. To make it more likely that the stationarity assumption holds, the variance of the random effects is set to a lower value than in the MAR.RE setting.

For each of the 10800 resulting datasets, we investigate the predictive accuracy of the following six models: AR, VAR, MAR.FE, MAR.RE, MVAR.FE, and MVAR.RE. Predictive accuracy performance measures are estimated using 10-blocked CV. All analyses are performed in R version 4.0.3 (R Core Team, 2020) and Julia version 1.5.3 (Bezanson et al., 2017). Multilevel models were estimated using the MixedModels.jl package version v3.1.4 (Bates et al., 2016). In the OSF page of the project, we include the simulation scripts to run the analyses reported in this article using R (see https://osf.io/rs6un/).

4.2. *Results*

In what follows, we discuss the simulation results for every data generating model consecutively. The average obtained MSPE values for the data generated from person-specific VAR(1) and multilevel VAR(1) models with fixed and random effects are shown in Table 1,[3] whereas Table 2 displays the results for the AR, MAR.FE, and MAR.RE data, for which the number of variables always equals 4.

Person-Specific VAR(1) Table 1 and Fig. 1 show that MVAR.RE exhibits the best predictive accuracy across all levels of the number of variables, persons and measurement occasions within persons. We also observe that the MSPE values obtained when fitting VAR and MVAR.RE increase with the number of variables and decrease with the number of measurement occasions within persons. Interestingly, we observe that when fitting VAR, predictive accuracy improves

[3] For each of the data generating models, we conducted a mixed ANOVA to investigate how the MSPE values of the six proposed methods are affected by the number of variables, the number of measurement occasions within persons, and the number of persons. Results show that there are large main effects of the number of variables and the estimation model. These main effects are qualified by interactions between the number of variables and the estimation model, and the number of variables and the number of measurement occasions. The results are included in the supplementary material.

Table 1.
Simulation results of study I. The MSPE estimates (standard deviation in parentheses) when data are generated using person-specific and multilevel VAR(1) models.

Population model	Method	Number of variables			Number of persons			Measurement occasions		
		$P=2$	$P=4$	$P=8$	$N=20$	$N=60$	$N=120$	$T=50$	$T=100$	$T=200$
VAR	AR	1.050	1.135	1.358	1.179	1.182	1.182	1.175	1.182	1.186
		(0.026)	(0.019)	(0.042)	(0.133)	(0.133)	(0.135)	(0.104)	(0.135)	(0.156)
	MAR.FE	1.039	1.132	1.358	1.173	1.178	1.178	1.157	1.180	1.193
		(0.021)	(0.018)	(0.044)	(0.138)	(0.136)	(0.138)	(0.115)	(0.138)	(0.154)
	MAR.RE	1.030	1.115	1.337	1.160	1.161	1.160	1.141	1.164	1.177
		(0.021)	(0.016)	(0.049)	(0.133)	(0.132)	(0.134)	(0.105)	(0.133)	(0.154)
	MVAR.FE	1.020	1.033	1.065	1.039	1.040	1.040	1.033	1.039	1.046
		(0.020)	(0.014)	(0.021)	(0.031)	(0.023)	(0.025)	(0.023)	(0.023)	(0.031)
	MVAR.RE	**1.011**	**1.018**	**1.025**	**1.023**	**1.017**	**1.014**	**1.025**	**1.018**	**1.012**
		(0.020)	(0.014)	(0.016)	(0.025)	(0.013)	(0.010)	(0.023)	(0.015)	(0.011)
	VAR	1.044	1.079	1.150	1.091	1.091	1.090	1.165	1.073	1.034
		(0.033)	(0.048)	(0.095)	(0.080)	(0.077)	(0.077)	(0.087)	(0.036)	(0.018)
MVAR.FE	AR	1.050	1.133	1.390	1.190	1.192	1.191	1.186	1.191	1.196
		(0.026)	(0.021)	(0.043)	(0.149)	(0.148)	(0.148)	(0.116)	(0.150)	(0.172)
	MAR.FE	1.023	1.103	1.381	1.168	1.169	1.168	1.150	1.171	1.185
		(0.019)	(0.016)	(0.045)	(0.157)	(0.156)	(0.156)	(0.132)	(0.158)	(0.175)
	MAR.RE	1.023	1.103	1.366	1.165	1.164	1.163	1.143	1.167	1.182
		(0.020)	(0.016)	(0.050)	(0.151)	(0.149)	(0.150)	(0.122)	(0.151)	(0.171)
	MVAR.FE	**1.001**	**1.002**	**1.002**	**1.004**	**1.002**	**1.001**	**1.003**	**1.002**	**1.001**
		(0.019)	(0.013)	(0.011)	(0.021)	(0.011)	(0.009)	(0.020)	(0.013)	(0.009)
	MVAR.RE	1.002	1.004	1.009	1.010	1.004	1.002	1.008	1.005	1.002
		(0.019)	(0.013)	(0.013)	(0.023)	(0.012)	(0.009)	(0.021)	(0.014)	(0.010)
	VAR	1.043	1.079	1.147	1.090	1.090	1.089	1.163	1.072	1.034
		(0.033)	(0.048)	(0.093)	(0.079)	(0.076)	(0.076)	(0.085)	(0.036)	(0.018)
MVAR.RE	AR	1.080	1.223	1.498	1.092	1.092	1.093	1.275	1.266	1.260
		(0.028)	(0.026)	(0.033)	(0.084)	(0.081)	(0.081)	(0.166)	(0.178)	(0.184)
	MAR.FE	1.096	1.279	1.608	1.324	1.328	1.331	1.304	1.330	1.349
		(0.029)	(0.039)	(0.059)	(0.218)	(0.216)	(0.216)	(0.198)	(0.217)	(0.230)
	MAR.RE	1.062	1.206	1.482	1.251	1.249	1.250	1.246	1.251	1.253
		(0.023)	(0.024)	(0.034)	(0.179)	(0.175)	(0.175)	(0.168)	(0.178)	(0.183)
	MVAR.FE	1.071	1.179	1.456	1.223	1.239	1.244	1.219	1.237	1.250
		(0.026)	(0.028)	(0.053)	(0.158)	(0.169)	(0.172)	(0.152)	(0.167)	(0.178)
	MVAR.RE	**1.019**	**1.040**	**1.090**	**1.055**	**1.048**	**1.046**	**1.077**	**1.046**	**1.025**
		(0.021)	(0.022)	(0.042)	(0.050)	(0.040)	(0.035)	(0.052)	(0.031)	(0.018)
	VAR	1.043	1.077	1.158	1.267	1.267	1.268	1.168	1.074	1.035
		(0.032)	(0.048)	(0.100)	(0.178)	(0.175)	(0.175)	(0.094)	(0.040)	(0.020)

Values in bold denote the minimum MSPE estimate

considerably when the number of measurement occasions increases. However, this improvement is not sufficient to outperform MVAR.RE. Table 3 shows how often each estimation model was selected as the best model, using the minimum MSPE rule and the one standard error rule (Bulteel et al. 2018a; Friedman et al. 2001). Using the latter rule, a more parsimonious model is selected if the MSPE falls within one standard deviation of the minimum. This rule is often applied to compare models with different numbers of parameters. We observe that when data are generated using VAR, MVAR.RE yields the lowest prediction error for 95.3% of the generated datasets; also the one standard error rule almost always favors MVAR.RE.

Table 2.

Simulation results of study I. The MSPE estimates (standard deviation in parentheses) when data are generated using person-specific and multilevel AR(1) models and the number of variables is four.

Population model	Method	Number of persons			Measurement occasions			Model selection	
		$N = 20$	$N = 60$	$N = 120$	$T = 50$	$T = 100$	$T = 200$	MSPE	One std. error
AR	AR	1.028	1.028	1.028	1.048	1.024	1.012	0.000	0.000
		(0.024)	(0.019)	(0.017)	(0.018)	(0.012)	(0.009)		
	MAR.FE	1.017	1.018	1.017	1.018	1.018	1.017	0.019	0.020
		(0.018)	(0.011)	(0.009)	(0.017)	(0.012)	(0.009)		
	MAR.RE	**1.010**	**1.009**	**1.008**	**1.012**	**1.009**	**1.005**	0.979	0.978
		(0.018)	(0.011)	(0.009)	(0.017)	(0.012)	(0.009)		
	MVAR.FE	1.019	1.019	1.018	1.020	1.019	1.017	0.000	0.000
		(0.018)	(0.011)	(0.009)	(0.018)	(0.012)	(0.009)		
	MVAR.RE	1.016	1.011	1.009	1.018	1.012	1.007	0.002	0.002
		(0.019)	(0.012)	(0.009)	(0.018)	(0.012)	(0.009)		
	VAR	1.081	1.082	1.081	1.147	1.066	1.031	0.000	0.000
		(0.052)	(0.052)	(0.050)	(0.022)	(0.014)	(0.009)		
MAR.FE	AR	1.028	1.028	1.028	1.048	1.024	1.012	0.000	0.000
		(0.024)	(0.020)	(0.018)	(0.019)	(0.011)	(0.009)		
	MAR.FE	**1.002**	**1.001**	**1.001**	**1.002**	**1.001**	**1.001**	0.833	0.819
		(0.018)	(0.011)	(0.009)	(0.018)	(0.011)	(0.009)		
	MAR.RE	**1.002**	1.002	**1.001**	1.002	1.002	**1.001**	0.128	0.140
		(0.018)	(0.011)	(0.009)	(0.018)	(0.011)	(0.009)		
	MVAR.FE	1.004	1.002	**1.001**	1.003	1.002	**1.001**	0.038	0.041
		(0.018)	(0.011)	(0.009)	(0.018)	(0.011)	(0.009)		
	MVAR.RE	1.008	1.004	1.002	1.008	1.005	1.002	0.001	0.000
		(0.018)	(0.011)	(0.009)	(0.018)	(0.011)	(0.009)		
	VAR	1.080	1.080	1.079	1.144	1.064	1.030	0.000	0.000
		(0.051)	(0.051)	(0.049)	(0.023)	(0.012)	(0.010)		
MVAR.RE	AR	1.026	1.025	1.026	1.043	1.023	1.011	0.000	0.000
		(0.025)	(0.018)	(0.016)	(0.021)	(0.013)	(0.009)		
	MAR.FE	1.150	1.162	1.166	1.142	1.162	1.173	0.000	0.000
		(0.045)	(0.033)	(0.027)	(0.031)	(0.032)	(0.038)		
	MAR.RE	**1.013**	**1.011**	**1.012**	**1.019**	**1.011**	**1.006**	0.996	0.994
		(0.022)	(0.013)	(0.010)	(0.020)	(0.013)	(0.009)		
	MVAR.FE	1.151	1.162	1.165	1.144	1.163	1.172	0.000	0.000
		(0.044)	(0.032)	(0.026)	(0.031)	(0.031)	(0.037)		
	MVAR.RE	1.020	1.014	1.013	1.025	1.014	1.007	0.004	0.006
		(0.023)	(0.014)	(0.010)	(0.021)	(0.013)	(0.009)		
	VAR	1.078	1.077	1.078	1.139	1.064	1.030	0.000	0.000
		(0.051)	(0.048)	(0.047)	(0.023)	(0.014)	(0.010)		

Values in bold denote the minimum MSPE estimate

Multilevel VAR(1) with Fixed Effects Simulation results show that fitting the correct model (i.e., MVAR.FE) yields the best overall predictive performance. The MVAR.FE results follow the expected trends, in that the MSPE values improve with increasing numbers of persons and measurement accessions, and a decreasing number of variables. We also observe that MVAR.RE performs only slightly worse than the correct model in the following settings: when the number of variables is 2, the number of persons is 120, and the number of time points is 200 (see Fig. 2).

Multilevel VAR(1) with Random Effects When using MVAR.RE as a data generating model, the best predictions are obtained when fitting the MVAR.RE model. Interestingly, the person-

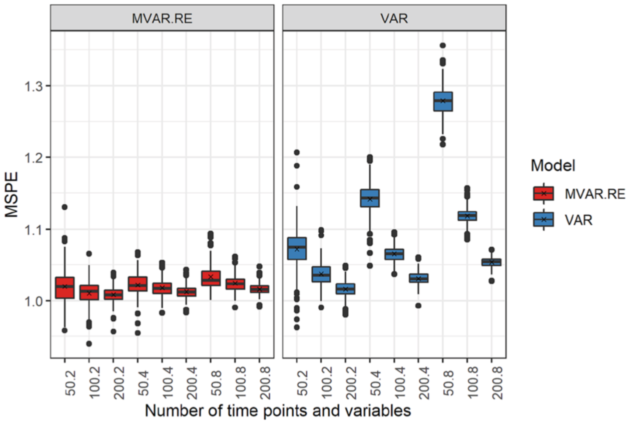

FIGURE 1.
Simulation results of the VAR setting of study I. Distribution of MSPE estimates for VAR and MVAR.RE. The boxplots represent the distribution of the MSPE values for each combination of the number of time points (i.e., 50, 100, and 200) and number of variables (i.e., 2, 4, and 8).

TABLE 3.
Simulation results of study I. The proportion of simulated person-specific and multilevel VAR(1) data sets for which a particular model was selected using the minimum MSPE rule and the one standard error rule.

Population model	Model Selection		
	Method	MSPE	One std. error
MVAR.FE	AR	0.000	0.000
	MAR.FE	0.000	0.000
	MAR.RE	0.000	0.000
	MVAR.FE	0.953	0.954
	MVAR.RE	0.047	0.046
	VAR	0.000	0.000
MVAR.RE	AR	0.000	0.000
	MAR.FE	0.000	0.000
	MAR.RE	0.000	0.000
	MVAR.FE	0.000	0.001
	MVAR.RE	1.000	0.999
	VAR	0.000	0.000
VAR	AR	0.000	0.000
	MAR.FE	0.000	0.000
	MAR.RE	0.000	0.000
	MVAR.FE	0.054	0.051
	MVAR.RE	0.946	0.949
	VAR	0.000	0.000

specific VAR(1) model has the second-best predictive accuracy. This can be better observed in Fig. 3, revealing that increasing the number of measurement occasions improves the performance of person-specific VAR(1) models, with this improvement being more pronounced when the number of variables increases. We also observe that FE models perform quite badly, since they cannot handle individual differences.

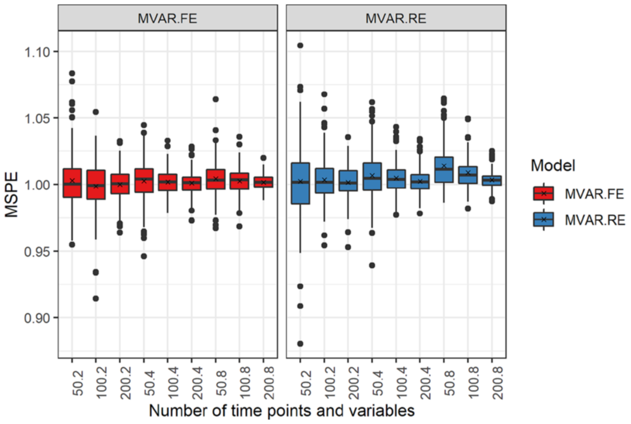

FIGURE 2.
Simulation results of the MVAR.FE setting of study I. Distribution of MSPE estimates for MVAR.FE and MVAR.RE. The boxplots represent the distribution of the MSPE values for each combination of the number of time points (i.e., 50, 100, and 200) and number of variables (i.e., 2, 4, and 8).

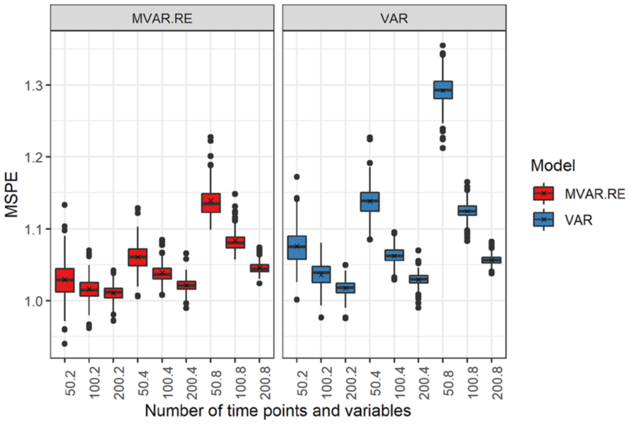

FIGURE 3.
Simulation results of the MVAR.RE setting of study I. Distribution of MSPE estimates for MVAR.RE and VAR. The boxplots represent the distribution of the MSPE values for each combination of the number of time points (i.e., 50, 100, and 200) and number of variables (i.e., 2, 4, and 8).

Person-Specific and Multilevel AR(1) For data generated on the basis of a person-specific AR(1) model, we observe that MAR.RE has the best predictive accuracy across all levels of the number of persons and measurement occasions within persons. In line with the results for the data generated on the basis of person-specific VAR models, the predictive accuracy of AR improves however with the number of measurement occasions. For the MAR.FE case, we see that across all conditions using the corresponding estimation model (MAR.FE) yields the best overall predictive performance, which shows that the true model generalizes best to unseen data. However, MAR.RE and MVAR.FE have a similar predictive performance when the number of persons is 120 and the number of measurement occasions is 200. When the data generating process is a MAR.RE model, results show that estimating this model yields the smallest MSPE across all levels of the number of persons and measurement occasions within persons. Yet, the MSPE values slightly decrease with the number of measurement occasions, and the number of persons. A similar but stronger pattern is found when fitting the MVAR.RE model.

In sum, the best prediction results are obtained when the estimation model matches the true model, except when the data are generated using person-specific AR(1) and VAR(1) models. Fitting these person-specific models to data very often leads to large MSPE values, showing that person-specific models cannot adequately capture the regularities of the time series under study in most settings. Yet, as predicted, person-specific models perform slightly better when the number of measurement occasions increases and the number of variables decreases; the number of persons seems to matter less.

5. Study II: Evaluating the Effect of Contemporaneous Multicollinearity

We conducted a second simulation study to investigate the effect of contemporaneous multicollinearity, caused by using multiple indicators of a few latent constructs. This setting makes sense as the concrete items used in intensive longitudinal studies are often motivated this way (e.g., Pe et al. 2015; Sels et al. 2016). Based on Bulteel et al. (2018b), we predict that using a model that takes the latent structure into account, such as the dimension reduction-based approach proposed in Sect. 2.3, will yield a better predictive accuracy than applying multilevel VAR(1) models on the original variables.

5.1. Simulation Design

In this study, we follow Bulteel et al. (2018b) and we fix the number of variables to six and the number of individuals to 60. We manipulate (i) the number of measurement occasions per person: 50 and 100, (ii) the amount of measurement error on the data: 5% and 50%, and (iii) the data generating model, including different amounts of collinearity. Specifically, we use six data generating models which all impose a latent structure and thus contemporaneous correlations on the variables: MAR.FE with one component, MAR.RE with one component, MVAR.FE with 2 or 3 components, and MVAR.RE with 2 or 3 components. We again include 100 replicates per design cell, yielding 2400 simulated datasets. Unlike Study I, we did not fix the true effects across the replicates within a cell, since we do not consider estimation accuracy here.

In the MAR cases, all six variables have a loading of one on a single component. The scores on this component follow an AR(1) process. The fixed autoregressive effects of the MAR.FE and MAR.RE settings are sampled from a Uniform distribution on the interval $[0.2, 0.6]$. In the MAR.RE setting, the random effects are drawn from a Gaussian distribution of which the mean corresponds to the fixed autoregressive effects, while each of the random effect variances amounted to 0.1 and the covariances to zero. The standard deviation of the within-individual innovations σ_ϵ was set to one for all individuals.

In the VAR cases, the variables exhibit either a two or a three component structure, whereas the component scores adhere to a VAR(1) process. In the two component conditions, the first four variables have a loading of one on the first component and the rest of the variables on the second component. In the three component conditions, the first two variables have a loading of one on the first component, the third and fourth variables on the second component, and the last two variables on the third component. The fixed autoregressive effects are again drawn from a uniform distribution on the interval $[0.2, 0.6]$. The fixed cross-regressive effects are sampled from a uniform distribution on the interval $[0.05, 0.20]$. To guarantee that the generated time series are stationary, the regression weights are multiplied by $0.99/(|\lambda_p|)$, where $|\lambda_p|$ denotes the absolute value of the maximum eigenvalue of the transition matrix. Furthermore, the diagonal elements of the covariance matrix of the within-individual innovations $\mathbf{\Sigma}_\epsilon$ are set to one, and the off-diagonal elements to 0.2. In the MVAR.RE scenario, the fixed autoregressive and cross-regressive effects are generated as was done in the MVAR.FE scenario. The random effects are generated from a multivariate Gaussian distribution for which the means equal the fixed autoregressive and cross-regressive effects, and with a diagonal covariance matrix in which the variances are set to 0.025. We again ensured that the generated time series of each individual conformed to the stationary assumption by checking the eigenvalues of the individual transition matrices.

Finally, we add measurement error to each simulated dataset, which was drawn from a standard normal distribution. This measurement noise was rescaled such that it was expected to equal either 5% or 50% of the total variance in the dataset.

Before evaluating predictive accuracy, we assessed the performance of the CHull procedure in retrieving the number of underlying components. Results show that this procedure adequately selects the true number of components. Specifically, when data contain 5% of measurement noise, CHull selects the true number of components in all 1200 datasets. When the measurement noise amounts to 50% and the regressive effects are random, CHull yields the correct amount of components in 597 out of the 600 datasets; for the fixed effects datasets, this holds for 580 out of the 600 datasets. Given this excellent result, we will firstly investigate the predictive accuracy when using the correct number of components. We have conducted additional analyses for the replicates in which CHull selects an incorrect number of components. For all these cases, CHull indicated the number of components to be one while the true number of components is three. The results are included in the supplementary material, and they follow the same pattern as to when the true number of components is used.

To evaluate predictive accuracy, we again performed 10-block CV. For each simulated dataset, we compare the predictive accuracy when applying the six estimation models from Study 1 on the raw dataset (i.e., using the original variables) as well as on the transformed data (i.e., after dimension reduction, denoted by adding PC. in the method label). To make the MSPE values of the analyses on the raw and transformed data comparable, we rescaled the prediction errors of the 'raw' analyses by dividing them by the standard deviation of the corresponding variable in the training data.

5.2. Results

The results for the MAR.FE and MAR.RE cases, in which the six variables load on a single underlying component, are shown in Table 4. Note that we do not present PC.VAR and PC.MVAR results, since after the dimension reduction step with $Q = 1$, only a single variable remains. Interestingly, simulation results show that running MVAR.FE on the original variables (hence, without dimension reduction) yields the best predictive accuracy results. Importantly, we note that in line with the findings in Bulteel et al. (2018b), applying the dimension reduction procedure with $Q = 1$ yields better predictive accuracy than estimating person-specific AR(1) and VAR(1) models. Not surprisingly and in line with Study I, both PC.MAR.FE and PC.MAR.RE perform almost equally well across all conditions. Moreover, results show that for all the methods, predictive accuracy improves when the number of measurement occasions per person increases. We also note that the MSPE values cannot be compared across the different levels of measurement noise because these values are influenced by the variance of the original variables, which increases with measurement error. For the same reason, the results in Study II cannot be compared with that of Study I. In terms of model selection, Table 4 includes the proportion of datasets for which a particular model was selected using the minimum MSPE rule and the one standard error rule. The two criteria favor MVAR.FE, even when effects are considered to be random.

Tables 5 and 6 present the results for the settings in which the latent structure is determined by two and three components, respectively. We again observe that multilevel models applied on the original variables show the best predictive performance: MVAR.FE shows the best predictive

TABLE 4.
Simulation results of study II. The MSPE estimates (standard deviation in parentheses) when the data generation model is a multilevel AR(1) model with one component.

	Fixed effects						Random effects					
	Error 5%		Error 50%		Model selection		Error 5%		Error 50%		Model selection	
Model	$T = 50$	$T = 100$	$T = 50$	$T = 100$	MSPE	One std. error	$T = 50$	$T = 100$	$T = 50$	$T = 100$	MSPE	One std. error
PC.AR	0.880	0.854	0.955	0.932	0.000	0.000	0.881	0.860	0.950	0.936	0.000	0.000
	(0.091)	(0.085)	(0.054)	(0.051)			(0.086)	(0.090)	(0.051)	(0.052)		
PC.MAR.FE	0.860	0.844	0.940	0.925	0.000	0.000	0.860	0.851	0.936	0.929	0.000	0.000
	(0.090)	(0.085)	(0.054)	(0.051)			(0.085)	(0.089)	(0.050)	(0.052)		
PC.MAR.RE	0.860	0.844	0.941	0.925	0.000	0.000	0.861	0.851	0.937	0.929	0.000	0.000
	(0.090)	(0.085)	(0.054)	(0.051)			(0.085)	(0.089)	(0.050)	(0.052)		
AR	0.888	0.860	0.997	0.969	0.000	0.000	0.888	0.866	0.992	0.972	0.000	0.000
	(0.089)	(0.083)	(0.037)	(0.034)			(0.084)	(0.087)	(0.034)	(0.035)		
MAR.FE	0.850	0.841	0.951	0.947	0.025	0.025	0.849	0.847	0.948	0.949	0.025	0.033
	(0.085)	(0.081)	(0.036)	(0.035)			(0.079)	(0.085)	(0.033)	(0.035)		
MAR.RE	0.850	0.841	0.952	0.948	0.005	0.005	0.850	0.847	0.949	0.949	0.005	0.005
	(0.085)	(0.081)	(0.036)	(0.035)			(0.079)	(0.085)	(0.033)	(0.035)		
MVAR.FE	**0.846**	**0.836**	**0.923**	**0.917**	0.953	0.955	**0.845**	**0.843**	**0.920**	**0.920**	0.948	0.950
	(0.087)	(0.083)	(0.052)	(0.051)			(0.081)	(0.088)	(0.049)	(0.051)		
MVAR.RE	0.852	0.839	0.929	0.920	0.018	0.015	0.851	0.845	0.926	0.923	0.023	0.013
	(0.088)	(0.084)	(0.052)	(0.051)			(0.082)	(0.088)	(0.049)	(0.051)		
VAR	1.006	0.907	1.099	0.994	0.000	0.000	1.005	0.913	1.094	0.998	0.000	0.000
	(0.096)	(0.096)	(0.057)	(0.056)			0.096	0.096	0.057	0.056		

Values in bold denote the minimum MSPE estimate
The model selection columns indicate how often a particular model was selected using the minimum MSPE rule and the one standard error rule

TABLE 5.
Simulation results of study II. The MSPE estimates (standard deviation in parentheses) when the data generation model is a multilevel VAR(1) model with two components

	Fixed effects						Random effects					
	Error 5%		Error 50%		Model selection		Error 5%		Error 50%		Model selection	
Model	$T = 50$	$T = 100$	$T = 50$	$T = 100$	MSPE	One std. error	$T = 50$	$T = 100$	$T = 50$	$T = 100$	MSPE	One std. error
PC.VAR	0.854	0.822	0.951	0.921	0.000	0.000	0.807	0.782	0.903	0.867	0.000	0.000
	(0.082)	(0.078)	(0.050)	(0.046)			(0.107)	(0.101)	(0.079)	(0.092)		
PC.MAR.FE	0.807	0.800	0.925	0.911	0.000	0.000	0.750	0.748	0.877	0.858	0.000	0.000
	(0.080)	(0.078)	(0.047)	(0.044)			(0.112)	(0.105)	(0.075)	(0.093)		
PC.MAR.RE	0.807	0.801	0.925	0.911	0.000	0.000	0.735	0.727	0.869	0.845	0.000	0.000
	(0.080)	(0.078)	(0.047)	(0.044)			(0.113)	(0.106)	(0.078)	(0.095)		
PC.MVAR.FE	0.804	0.797	0.915	0.903	0.000	0.000	0.746	0.743	0.867	0.852	0.000	0.000
	(0.079)	(0.077)	(0.049)	(0.046)			(0.112)	(0.105)	(0.076)	(0.094)		
PC.MVAR.RE	0.805	0.797	0.916	0.904	0.000	0.000	0.721	0.707	0.854	0.831	0.003	0.010
	(0.079)	(0.077)	(0.049)	(0.046)			(0.113)	(0.106)	(0.079)	(0.096)		
AR	0.847	0.828	0.977	0.955	0.000	0.000	0.759	0.742	0.915	0.885	0.000	0.000
	(0.078)	(0.074)	(0.038)	(0.032)			(0.114)	(0.106)	(0.071)	(0.090)		
MAR.FE	0.811	0.808	0.934	0.931	0.000	0.000	0.755	0.757	0.895	0.887	0.000	0.000
	(0.075)	(0.073)	(0.035)	(0.032)			(0.108)	(0.102)	(0.061)	(0.082)		
MAR.RE	0.811	0.809	0.935	0.932	0.000	0.000	0.737	0.732	0.885	0.872	0.000	0.000
	(0.075)	(0.073)	(0.035)	(0.032)			(0.110)	(0.104)	(0.066)	(0.088)		
MVAR.FE	**0.792**	**0.789**	**0.899**	**0.894**	0.998	0.995	0.737	0.737	0.854	0.844	0.018	0.020
	(0.077)	(0.076)	(0.047)	(0.046)			(0.109)	(0.104)	(0.072)	(0.092)		
MVAR.RE	0.796	0.792	0.905	0.897	0.003	0.005	**0.713**	**0.702**	**0.846**	**0.827**	0.980	0.970
	(0.078)	(0.076)	(0.047)	(0.046)			(0.111)	(0.105)	(0.077)	(0.096)		
VAR	0.944	0.857	1.072	0.971	0.000	0.000	0.822	0.748	0.984	0.884	0.000	0.000
	(0.091)	(0.082)	(0.056)	(0.049)			(0.127)	(0.111)	(0.090)	(0.103)		

Values in bold denote the minimum MSPE estimate
The model selection columns indicate how often a particular model was selected using the minimum MSPE rule and the one standard error rule

TABLE 6.
Simulation results of study II. The MSPE estimates (standard deviation in parentheses) when the data generation model is a multilevel VAR(1) model with three components .

	Fixed effects						Random effects					
	Error 5%		Error 50%		Model selection		Error 5%		Error 50%		Model selection	
Model	$T = 50$	$T = 100$	$T = 50$	$T = 100$	MSPE	One std. error	$T = 50$	$T = 100$	$T = 50$	$T = 100$	MSPE	One std. error
PC.VAR	0.819	0.750	0.940	0.883	0.000	0.000	0.765	0.742	0.871	0.821	0.000	0.000
	(0.078)	(0.091)	(0.050)	(0.063)			(0.113)	(0.117)	(0.089)	(0.097)		
PC.MAR.FE	0.792	0.767	0.913	0.892	0.000	0.000	0.648	0.655	0.815	0.800	0.000	0.000
	(0.075)	(0.082)	(0.044)	(0.052)			(0.122)	(0.132)	(0.092)	(0.103)		
PC.MAR.RE	0.793	0.767	0.914	0.892	0.000	0.000	0.631	0.632	0.803	0.779	0.000	0.000
	(0.075)	(0.082)	(0.044)	(0.051)			(0.122)	(0.132)	(0.095)	(0.106)		
PC.MVAR.FE	0.740	0.712	0.879	0.853	0.000	0.000	0.609	0.612	0.783	0.764	0.000	0.000
	(0.077)	(0.089)	(0.049)	(0.063)			(0.120)	(0.129)	(0.095)	(0.108)		
PC.MVAR.RE	0.742	0.713	0.880	0.854	0.000	0.000	0.574	0.563	0.764	0.733	0.000	0.000
	(0.078)	(0.089)	(0.049)	(0.063)			(0.119)	(0.125)	(0.097)	(0.111)		
AR	0.801	0.765	0.953	0.921	0.000	0.000	0.626	0.620	0.832	0.800	0.000	0.000
	(0.079)	(0.087)	(0.041)	(0.050)			(0.124)	(0.133)	(0.095)	(0.107)		
MAR.FE	0.767	0.746	0.911	0.898	0.000	0.000	0.632	0.641	0.823	0.814	0.000	0.000
	(0.074)	(0.086)	(0.037)	(0.049)			(0.120)	(0.131)	(0.085)	(0.097)		
MAR.RE	0.767	0.747	0.912	0.899	0.000	0.000	0.611	0.613	0.808	0.789	0.000	0.000
	(0.074)	(0.085)	(0.038)	(0.049)			(0.120)	(0.131)	(0.090)	(0.104)		
MVAR.FE	**0.729**	**0.706**	**0.864**	**0.845**	0.998	0.995	0.603	0.608	0.774	0.758	0.000	0.000
	(0.075)	(0.088)	(0.047)	(0.062)			(0.117)	(0.092)	(0.127)	(0.106)		
MVAR.RE	0.733	0.708	0.870	0.847	0.003	0.005	**0.567**	**0.559**	**0.757**	**0.729**	1.000	1.000
	(0.076)	(0.088)	(0.048)	(0.062)			(0.117)	(0.124)	(0.096)	(0.110)		
VAR	0.871	0.767	1.032	0.919	0.000	0.000	0.647	0.591	0.875	0.775	0.000	0.000
	(0.088)	(0.095)	(0.058)	(0.067)			(0.133)	(0.131)	(0.113)	(0.117)		

Values in bold denote the minimum MSPE estimate
The model selection columns indicate how often a particular model was selected using the minimum MSPE rule and the one standard error rule.

accuracy when effects are considered fixed, whereas MVAR.RE outperforms the other methods when effects are considered to be random. We also derive that the manipulated data characteristics impact performance as expected, with more measurement occasions implying higher predictive accuracy. It is interesting to note that when data are generated by imposing an MVAR.RE on the components, PC.MVAR.RE exhibits the second-best performance.

Hence, when using multiple indicator variables induces contemporaneous multicollinearity, applying dimension reduction techniques before running analyses improves the obtained predictive accuracy of person-specific AR(1) and VAR(1) models. Yet, simulation results show that analyzing the original data using multilevel models yields even better predictive accuracy than the methods that apply dimension reduction techniques.

6. Study III: Evaluating the Effect of Qualitative Individual Differences

In the multilevel random effect models, the individual differences in each regression coefficient are assumed to be Gaussian distributed. However, building on the examples provided by Bulteel et al. (2016a) and Ernst et al. (2019), it makes sense to assume that these differences in some cases are categorical and thus have a multimodal shape. Therefore, we ran a third simulation study to understand how the predictive accuracy of person-specific and multilevel VAR(1) models is affected by categorical individual differences in the autoregressive and cross-regressive effects. To this end, we simulated data that include clusters of individuals that share similar dynamics and analyzed them with the same six models that were applied in Study I. Although, among others, Bulteel et al. (2016a) and Ernst et al. (2019) proposed clustering approaches to VAR(1) modeling, we did not apply these techniques, because they are computationally intensive (i.e., to avoid local minima problems, multiple start procedures have to be run). Moreover, the implementation of K-blocked cross-validation is not so straightforward.

6.1. Simulation Design

To simulate data we build on the work of Bulteel et al. (2016a) and Ernst et al. (2019). Specifically, setting the number of variables to four, we manipulate five data characteristics: (i) the number of persons: 20 and 60; (ii) the number of measurement occasions per person: 50, 100, and 200; (iii) the number of clusters: 2 and 4; (iv) the cluster sizes: equal cluster size, minority case with one cluster including 10% of the persons only, and a majority case with one cluster including 60% of the persons (in the minority and majority cases, we assume that the rest of the persons are equally distributed across the other clusters), (v) the data generating model per cluster: MVAR.FE and MVAR.RE. Within each cell of the design, we generate 100 datasets, yielding 7200 datasets in total.

When generating these datasets, we start by simulating a different transition matrix for each cluster, imposing that the between-cluster differences pertain to the auto- and cross-regressive effects only. First, the cluster-specific autoregressive fixed effects are drawn from a uniform distribution on the interval [0.2, 0.60]. Within each cluster, the fixed cross-regressive effects are split in two equally large subsets; note that the split differs across the clusters. For each cluster separately, one subset is sampled from a uniform distribution on the interval [0.05, 0.20], and the other from a uniform distribution [-0.10, -0.05]. We again guaranteed stationarity, by accounting for the absolute value of the maximum eigenvalue of the matrix with fixed effects (see Study I). In the MVAR.RE settings, the random effects are simulated from a multivariate Gaussian distribution with the means given by the fixed autoregressive and cross-regressive effects of the cluster to which the person belongs, and a diagonal covariance matrix in which the variances were set to 0.025. Also here, we ensure that the generated time series are stationary by checking the eigenvalues of

Table 7.
Simulation results of study III. The MSPE (standard deviation in parentheses) estimates when the data generation model boils down to cluster-specific multilevel VAR(1) models with fixed effects.

Cluster size		$N = 20$ and 2 Clusters			$N = 60$ and 2 Clusters			$N = 20$ and 4 Clusters			$N = 60$ and 4 Clusters		
	Method	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
Equal	AR	1.094	1.063	1.052	1.090	1.069	1.055	1.091	1.068	1.053	1.094	1.066	1.055
		(0.032)	(0.021)	(0.017)	(0.020)	(0.014)	(0.011)	(0.033)	(0.023)	(0.014)	(0.018)	(0.012)	(0.011)
	MAR.FE	1.058	1.050	1.052	1.051	1.054	1.053	1.059	1.059	1.056	1.059	1.056	1.058
		(0.030)	(0.021)	(0.018)	(0.021)	(0.016)	(0.012)	(0.032)	(0.023)	(0.016)	(0.019)	(0.013)	(0.012)
	MAR.RE	1.055	1.045	1.044	1.048	1.049	1.046	1.054	1.051	1.046	1.054	1.048	1.047
		(0.029)	(0.020)	(0.017)	(0.020)	(0.013)	(0.011)	(0.031)	(0.022)	(0.014)	(0.018)	(0.012)	(0.010)
	MVAR.FE	1.040	1.032	1.030	1.029	1.032	1.030	1.051	1.049	1.045	1.049	1.044	1.046
		(0.028)	(0.022)	(0.017)	(0.019)	(0.013)	(0.013)	(0.030)	(0.022)	(0.014)	(0.017)	(0.013)	(0.011)
	MVAR.RE	**1.032**	**1.013**	**1.007**	**1.017**	**1.012**	**1.006**	**1.039**	**1.026**	**1.014**	**1.032**	**1.018**	**1.012**
		(0.028)	(0.018)	(0.014)	(0.018)	(0.010)	(0.007)	(0.029)	(0.020)	(0.012)	(0.015)	(0.011)	(0.008)
	VAR	1.151	1.063	1.029	1.146	1.067	1.031	1.147	1.067	1.030	1.149	1.065	1.031
		(0.034)	(0.019)	(0.014)	(0.020)	(0.012)	(0.007)	(0.035)	(0.022)	(0.012)	(0.018)	(0.012)	(0.009)
Minority	AR	1.089	1.070	1.055	1.090	1.066	1.054	1.093	1.068	1.054	1.088	1.065	1.054
		(0.031)	(0.023)	(0.016)	(0.020)	(0.014)	(0.013)	(0.034)	(0.021)	(0.017)	(0.019)	(0.012)	(0.009)
	MAR.FE	1.048	1.049	1.047	1.045	1.046	1.046	1.058	1.058	1.056	1.055	1.055	1.056
		(0.031)	(0.022)	(0.015)	(0.019)	(0.014)	(0.012)	(0.030)	(0.020)	(0.017)	(0.019)	(0.013)	(0.010)
	MAR.RE	1.048	1.049	1.045	1.045	1.044	1.044	1.055	1.051	1.047	1.049	1.047	1.046
		(0.030)	(0.022)	(0.015)	(0.019)	(0.014)	(0.012)	(0.030)	(0.020)	(0.016)	(0.018)	(0.012)	(0.009)
	MVAR.FE	**1.019**	**1.016**	1.013	1.012	1.011	1.012	1.048	1.046	1.044	1.043	1.042	1.043
		(0.028)	(0.018)	(0.014)	(0.016)	(0.011)	(0.007)	(0.030)	(0.020)	(0.017)	(0.018)	(0.014)	(0.009)
	MVAR.RE	1.023	**1.016**	**1.008**	**1.011**	**1.007**	**1.005**	**1.040**	**1.025**	**1.014**	**1.027**	**1.018**	**1.011**
		(0.028)	(0.018)	(0.013)	(0.016)	(0.010)	(0.007)	(0.030)	(0.019)	(0.014)	(0.016)	(0.011)	(0.007)
	VAR	1.148	1.070	1.031	1.145	1.065	1.031	1.150	1.068	1.031	1.144	1.066	1.030
		(0.032)	(0.019)	(0.014)	(0.021)	(0.010)	(0.007)	(0.037)	(0.020)	(0.014)	(0.018)	(0.012)	(0.007)

TABLE 7.
continued

Cluster size	Method	$N = 20$ and 2 Clusters			$N = 60$ and 2 Clusters			$N = 20$ and 4 Clusters			$N = 60$ and 4 Clusters		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
Majority	AR	1.082	1.070	1.054	1.094	1.064	1.055	1.093	1.069	1.053	1.089	1.064	1.053
		(0.028)	(0.022)	(0.014)	(0.021)	(0.012)	(0.011)	(0.027)	(0.020)	(0.014)	(0.019)	(0.012)	(0.011)
	MAR.FE	1.045	1.055	1.052	1.054	1.049	1.053	1.058	1.058	1.053	1.054	1.052	1.054
		(0.028)	(0.022)	(0.015)	(0.020)	(0.014)	(0.013)	(0.027)	(0.021)	(0.016)	(0.018)	(0.013)	(0.012)
	MAR.RE	1.042	1.051	1.046	1.051	1.045	1.046	1.055	1.052	1.046	1.049	1.046	1.046
		(0.027)	(0.021)	(0.013)	(0.020)	(0.012)	(0.011)	(0.026)	(0.020)	(0.014)	(0.018)	(0.012)	(0.011)
	MVAR.FE	1.025	1.031	1.030	1.033	1.027	1.030	1.043	1.041	1.036	1.037	1.034	1.035
		(0.026)	(0.021)	(0.014)	(0.019)	(0.013)	(0.012)	(0.025)	(0.020)	(0.015)	(0.017)	(0.011)	(0.011)
	MVAR.RE	**1.019**	**1.017**	**1.009**	**1.020**	**1.009**	**1.007**	**1.036**	**1.024**	**1.013**	**1.025**	**1.016**	**1.010**
		(0.025)	(0.019)	(0.010)	(0.018)	(0.009)	(0.008)	(0.024)	(0.018)	(0.013)	(0.016)	(0.009)	(0.007)
	VAR	1.137	1.069	1.031	1.148	1.064	1.032	1.150	1.067	1.030	1.146	1.064	1.030
		(0.030)	(0.020)	(0.011)	(0.021)	(0.010)	(0.008)	(0.028)	(0.019)	(0.013)	(0.019)	(0.010)	(0.007)

Values in bold denote the minimum MSPE estimate

Table 8.

Simulation results of study III. The MSPE (standard deviation in parentheses) estimates when the data generation model boils down to cluster-specific multilevel VAR(1) models with random effects.

Cluster size	Method	$N = 20$ and 2 Clusters			$N = 60$ and 2 Clusters			$N = 20$ and 4 Clusters			$N = 60$ and 4 Clusters		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
Equal	AR	1.206	1.182	1.174	1.205	1.184	1.170	1.203	1.186	1.178	1.203	1.185	1.171
		(0.042)	(0.032)	(0.031)	(0.032)	(0.020)	(0.017)	(0.047)	(0.035)	(0.029)	(0.023)	(0.018)	(0.017)
	MAR.FE	1.216	1.225	1.234	1.223	1.232	1.236	1.225	1.237	1.244	1.224	1.242	1.248
		(0.045)	(0.047)	(0.048)	(0.043)	(0.029)	(0.032)	(0.058)	(0.046)	(0.047)	(0.028)	(0.031)	(0.036)
	MAR.RE	1.178	1.169	1.168	1.174	1.170	1.163	1.176	1.173	1.172	1.173	1.171	1.165
		(0.041)	(0.032)	(0.031)	(0.031)	(0.020)	(0.017)	(0.045)	(0.034)	(0.029)	(0.022)	(0.019)	(0.017)
	MVAR.FE	1.189	1.195	1.201	1.196	1.204	1.207	1.207	1.218	1.221	1.210	1.226	1.231
		(0.043)	(0.043)	(0.042)	(0.036)	(0.028)	(0.028)	(0.050)	(0.042)	(0.038)	(0.026)	(0.028)	(0.033)
	MVAR.RE	**1.078**	**1.044**	**1.023**	**1.064**	**1.038**	**1.021**	**1.079**	**1.045**	**1.025**	**1.069**	**1.042**	**1.021**
		(0.032)	(0.019)	(0.013)	(0.018)	(0.012)	(0.007)	(0.030)	(0.019)	(0.012)	(0.018)	(0.011)	(0.007)
	VAR	1.146	1.065	1.030	1.144	1.065	1.030	1.143	1.066	1.033	1.148	1.068	1.031
		(0.032)	(0.020)	(0.014)	(0.021)	(0.012)	(0.007)	(0.033)	(0.020)	(0.012)	(0.021)	(0.011)	(0.007)
Minority	AR	1.200	1.181	1.175	1.206	1.182	1.173	1.208	1.184	1.174	1.207	1.181	1.174
		(0.046)	(0.033)	(0.030)	(0.027)	(0.022)	(0.027)	(0.044)	(0.032)	(0.026)	(0.026)	(0.019)	(0.017)
	MAR.FE	1.207	1.214	1.229	1.218	1.222	1.230	1.229	1.232	1.241	1.233	1.235	1.249
		(0.050)	(0.045)	(0.046)	(0.035)	(0.032)	(0.040)	(0.053)	(0.042)	(0.047)	(0.036)	(0.030)	(0.034)
	MAR.RE	1.171	1.167	1.169	1.175	1.168	1.166	1.180	1.172	1.168	1.177	1.167	1.167
		(0.045)	(0.033)	(0.030)	(0.027)	(0.022)	(0.027)	(0.043)	(0.032)	(0.026)	(0.025)	(0.019)	(0.017)
	MVAR.FE	1.166	1.170	1.180	1.177	1.182	1.188	1.213	1.210	1.219	1.218	1.218	1.231
		(0.042)	(0.037)	(0.036)	(0.029)	(0.026)	(0.030)	(0.051)	(0.039)	(0.040)	(0.032)	(0.027)	(0.031)
	MVAR.RE	**1.075**	**1.044**	**1.023**	**1.064**	**1.038**	**1.021**	**1.084**	**1.045**	**1.024**	**1.071**	**1.039**	**1.021**
		(0.030)	(0.019)	(0.012)	(0.015)	(0.011)	(0.008)	(0.031)	(0.019)	(0.014)	(0.019)	(0.010)	(0.006)
	VAR	1.145	1.066	1.030	1.144	1.065	1.031	1.150	1.065	1.032	1.150	1.064	1.030
		(0.036)	(0.020)	(0.012)	(0.018)	(0.011)	(0.008)	(0.033)	(0.019)	(0.014)	(0.021)	(0.011)	(0.007)

Table 8.
continued

Cluster size	Method	$N = 20$ and 2 Clusters			$N = 60$ and 2 Clusters			$N = 20$ and 4 Clusters			$N = 60$ and 4 Clusters		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
Majority	AR	1.201	1.181	1.182	1.206	1.182	1.172	1.200	1.189	1.180	1.207	1.185	1.172
		(0.039)	(0.034)	(0.033)	(0.024)	(0.021)	(0.019)	(0.051)	(0.028)	(0.033)	(0.025)	(0.022)	(0.020)
	MAR.FE	1.213	1.219	1.248	1.222	1.226	1.237	1.218	1.236	1.244	1.224	1.236	1.241
		(0.047)	(0.045)	(0.050)	(0.029)	(0.033)	(0.035)	(0.059)	(0.046)	(0.055)	(0.036)	(0.033)	(0.042)
	MAR.RE	1.174	1.168	1.176	1.175	1.168	1.165	1.172	1.176	1.174	1.175	1.171	1.165
		(0.038)	(0.034)	(0.032)	(0.023)	(0.021)	(0.019)	(0.049)	(0.029)	(0.033)	(0.025)	(0.021)	(0.021)
	MVAR.FE	1.187	1.188	1.211	1.198	1.198	1.207	1.195	1.208	1.212	1.203	1.213	1.217
		(0.044)	(0.042)	(0.041)	(0.029)	(0.028)	(0.030)	(0.051)	(0.039)	(0.046)	(0.033)	(0.029)	(0.035)
	MVAR.RE	**1.077**	**1.046**	**1.026**	**1.069**	**1.040**	**1.021**	**1.075**	**1.048**	**1.025**	**1.069**	**1.040**	**1.021**
		(0.029)	(0.020)	(0.014)	(0.017)	(0.011)	(0.008)	(0.034)	(0.017)	(0.014)	(0.014)	(0.011)	(0.008)
	VAR	1.146	1.067	1.033	1.150	1.066	1.030	1.139	1.068	1.033	1.148	1.066	1.031
		(0.034)	(0.020)	(0.014)	(0.021)	(0.011)	(0.008)	(0.038)	(0.017)	(0.014)	(0.016)	(0.012)	(0.008)

Values in bold denote the minimum MSPE estimate

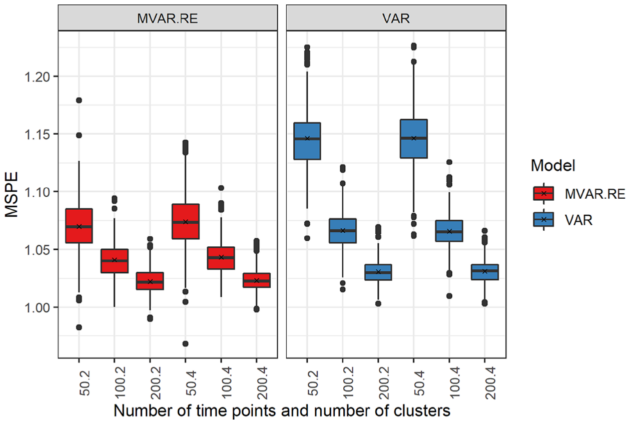

FIGURE 4.
Simulation results of the MVAR.RE setting of study III. Distribution of MSPE estimates for the method with the best and second-best performance across simulation conditions. The boxplots represent the distribution of the MSPE values for each combination of the number of time points (i.e., 50, 100, and 200) and clusters (i.e., 2 and 4).

the individual transition matrices. In the "Appendix", we included histograms of the distributions of the auto- and cross-regressive effects across persons when data are generated from a multilevel VAR(1) model with random effects, a person-specific VAR(1) model, or a cluster-specific VAR(1) model. These plots showcase the differences between the associated regressive effect distributions. For cluster-specific VAR(1) models with random effects, the obtained distributions deviate from a normal one, with the distribution shape being influenced by: (i) the differences between the fixed effects across clusters, (ii) the cluster size, and (iii) the differences between individuals that belong to the same cluster. Next, we evaluate the predictive accuracy of the six estimation models presented in Study I, by performing 10-block CV in each of the 7200 simulated datasets.

6.2. Results

Table 7 shows the predictive accuracy results for the settings that imply a different MVAR.FE within each cluster. The MVAR.RE model has the best overall predictive performance across all conditions, except for one condition. Specifically, when the number of persons is 20, the number of measurement occasions per person is 50, the number of clusters is 2, and one of the clusters includes two persons only (i.e., minority case), the MVAR.FE model yields the lowest MSPE. This makes sense as that condition implies that most of the persons (i.e., 48 participants) share the same transition matrix. Furthermore, we observe that the MSPE values of the models decrease when the number of persons and the number of measurement occasions per person increase. We note that when the number of clusters is 2, MVAR.FE shows the second-best performance. On the other hand, when the number of clusters is 4, VAR exhibits the second-best performance. This is no surprise since the different effects in the four clusters can be approximated less well by one set of fixed effects. Moreover, we note that the predictive performance of AR and VAR improves substantially when the number of measurement occasions per person increases. Regarding model selection, using the minimum MSPE implies that MVAR.RE is selected in 0.917% of the datasets, followed by MVAR.FE (0.081%), and MAR.RE (0.002%). A similar pattern is observed when the one standard error rule is used.

Table 8 presents the predictive accuracy results when the cluster-specific models are MVAR.RE models. Across all conditions, the best overall predictive performance is again obtained with MVAR.RE. Interestingly, the person-specific VAR(1) model now exhibits the second-best performance, and its MSPE diminishes when the number of measurement occasions grows larger (see Fig. 4). The results also show that the MSPE of MVAR.FE increases with the number of measurement occasions across all conditions, which might be explained by the shrinkage of the cluster-specific effects toward the overall mean. Finally, the model selection criteria pick MVAR.RE across all simulated datasets.

7. Discussion

Psychological research increasingly focuses on quantifying how complex processes evolve dynamically over time within individuals and on how this differs across individuals. Two popular approaches to model these dynamics are the person-specific and multilevel VAR(1) models, and their AR(1) counterparts. The application of these models is challenging, however, for several reasons: the amount of information available to base parameter estimation on is often limited, the variables under study can be highly contemporaneously correlated, and the individual differences may be categorical rather than Gaussian distributed. To assess whether the models under study adequately represent the dynamics of intensive longitudinal data and generalize well to unseen data, researchers may use predictive accuracy measures. These measures can also be used to perform model selection (e.g., Friedman et al. 2001). To study the effect of the challenging conditions mentioned above on the predictive accuracy of VAR(1) approaches, we conducted three simulation studies.

7.1. Simulation Results

In the first study, we investigated the effect of the number of measurement occasions per person, the number of persons, and the number of variables on the predictive accuracy of person-specific and multilevel VAR(1) models with fixed or random effects. By simulating data from six different model specifications, we showed that when data are generated using multilevel AR(1) and VAR(1) models, predictive accuracy tends to select the true data generating model even if the number of measurement occasions and the number of persons is low, and the number of variables is large. These results suggest that pooling information across persons by using multilevel techniques prevents overfitting when no individual differences are present or when individual differences are normally distributed. On the other hand, when data are generated from person-specific models, multilevel models with random effects generally yield better predictive accuracy. However, we note that the performance of person-specific AR(1) and VAR(1) considerably improves when the number of measurements occasions is large. These findings add to the results in Bulteel et al. (2018a), Krone et al. (2016), Krone et al. (2017), Krone et al. (2018), and Liu (2017). Additionally, we demonstrated that when within-individual dynamics are fixed across individuals, applying person-specific VAR(1) models is not a good idea since these models are too complex and likely to extract non-existing differences. However, when data were generated using a multilevel VAR(1) model with random effects, the person-specific VAR(1) model exhibited the second-best predictive performance. These results differ from the findings in Bulteel et al. (2018a), where person-specific VAR(1) models exhibit the worst predictive performance. This difference can be explained by the different simulation designs. In Bulteel et al. (2018a), simulations were based on real datasets. Therefore, when compared to Study I, there might be differences in the strength of the auto- and cross-regressive effects, and in the distribution of the random effects and within-persons innovations. Therefore, additional investigation is needed to assess the sensitivity of the MSPE to changes in the model parameters and distribution of the variance components.

We also investigated the effect of strong contemporaneous correlations between variables, due to an underlying latent construct structure. Here, we built on the proposal of Bulteel et al. (2018b). We presented a two-step approach to handle this multicollinearity, while taking into account the multilevel structure of intensive longitudinal datasets. In the first step, PCA is used to obtain a reduced dimensional representation of the predictors across all persons. In the second step, within-individual dynamics are quantified by specifying one of the VAR(1) models under study in the reduced space. In the second simulation study, we evaluated the predictive performance of this new approach. Simulation results showed that when multicollinearity is likely to be present in the data, applying the dimension reduction technique, and fitting person-specific or multilevel VAR(1) models in the reduced space yielded a better predictive accuracy than estimating person-specific VAR(1) models on the non-reduced data. These results are in line with the findings in Bulteel et al. (2018b). However, we note that in case of strong collinearity, multilevel VAR(1) models fitted on the non-reduced data exhibit better predictive performance than models fitted on the reduced data. These results are consistent with previous research on linear models. For example, Wainer (1976) stated that in case a set of predictor variables are perfectly correlated, any combination of them yields the same predictive accuracy. This result is somewhat worrisome, however, as the multilevel results on the original data may have a strongly different interpretation. Indeed, Bulteel et al. (2018b) showed that when variables are strongly correlated, the obtained estimates will reflect the well known bouncing beta problem, implying very large standard errors and estimation instability. Given that applying multilevel VAR(1) models to the reduced data and to the original variables returns estimates of the transition matrix that differ in matrix size, we did not study estimation accuracy when variables are strongly correlated. Consequently, further investigation is necessary to understand how an underlying latent structure affects the estimation of the transition matrix, as well as the bias and the variance trade off of the estimated target function. Finally, we note two limitations of the two-step approach. First, SCA does not remove the autocorrelation in the variables when person-centering the predictors, which can introduce biases in the model estimation step (Song and Zhang 2014). Second, we use orthogonal rotation because that renders the resulting components uncorrelated, implying that they do not have shared effects. However, whether using oblique rotation might yield better predictive performance remains an open question.

To shed light on the predictive performance of the proposed VAR(1) models when individual differences are to some extent categorical rather than Gaussian distributed, we conducted a third simulation study. In this study the simulated data included groups of individuals with similar dynamics. The simulation results showed that in general, the multilevel VAR(1) model with random effects exhibits better predictive performance than person-specific VAR(1) models, and multilevel VAR(1) models with fixed effects. However, the performance difference partly depends on the number of measurement occasions, the number of clusters, and the cluster sizes. Most importantly, the question remains whether person-specific VAR(1) models would in the end outperform the multilevel VAR(1) model with random effects, if the number of measurement occasions would still be strongly increased, although such settings would not occur very frequently in empirical practice.

7.2. *Directions for Future Research*

Although these three simulation studies highlighted how different data characteristics impact predictive accuracy, several questions remain unanswered. A first question pertains to our simulation settings. To keep computations feasible, we had to restrict the number of settings that we could investigate. Indeed, in simulation study III only, we already fitted 432000 (i.e., 7200 × 10 folds × 6 estimation models) different models. However, we are well aware that our conclusions do not necessarily generalize to other settings. For instance, it may very well be that increasing the differences between the clusters or the number of measurement occasions may lead to a better performance of the person-specific approaches.

A second question is how predictive accuracy behaves when one performs variable selection to target the analysis on a subset of relevant variables? In this regard, several approaches have been proposed to simultaneously perform variable selection and provide insight into within-individual dynamics. For example, lasso regression allows estimating person-specific VAR(1) models and performing variable selection, by setting some of the autoregressive and cross-regressive effects to zero (e.g., Bulteel et al. 2018a). In the context of multilevel models, some ℓ_1 regularized approaches have been proposed to perform variable selection (Müller et al. 2013). Finally, an alternative approach to model intensive longitudinal data is GIMME (Gates and Molenaar 2012) which estimates dynamic relations between variables, and simultaneously sets some of the coefficients to zero. Therefore, future research might focus on investigating the predictive accuracy of these methods as well.

A third question is related to how well factor analysis-based approaches handle multicollinearity between the variables and estimate dynamic relations. Browne and Nesselroade (2005) for instance proposed exploratory process factor analysis (EPFA). This approach assumes that within-individual processes are manifestations of underlying factors that follow a person-specific VAR(1) model. This model thus extracts a few latent factors while simultaneously modeling their dynamics. The implementation of cross-validation techniques for such factor analysis approaches is still under development (see, e.g., Bulteel et al. 2018b). Nevertheless, investigating the pros and cons of factor analysis-based and component analysis-based approaches seems worthwhile.

Recent studies have proposed extensions of person-specific VAR(1) models that allow quantifying categorical differences among individuals (e.g., Bulteel et al. 2016a; Ernst et al. 2019). These methods allow grouping individuals that share similar dynamical processes by applying clustering techniques to the autoregressive and cross-regressive effects in VAR(1) models. An area of further action is to investigate how cross-validation techniques can be applied to study the predictive accuracy of these clustering approaches.

Finally, in the simulation studies, we compared the predictive accuracy of competing explanations or models of the data. In practice, we would go on by choosing and interpreting the model that minimizes the prediction error, using the predictive accuracy measure in a relative sense. This leaves two sets of questions unanswered. First, to grasp better which model will win this predictive accuracy competition, we should shed further light on how predictive accuracy is related to the bias-variance trade-off in estimation accuracy. This would also help to use predictive accuracy in an absolute rather than relative way, by assessing whether any of the models under consideration actually implies low estimation bias and variance and thus reaches a good enough accuracy, and which data characteristics can be changed (i.e., more persons or measurement occasions) to achieve this absolute goal. Put differently, when analyzing empirical data, it can very well happen that all considered models are quite bad and then we would also want to detect that by looking at predictive accuracy. How to do this, remains an open question for now. A second question delves deeper into the relation between sample size planning, predictive and estimation accuracy, and power. In intensive longitudinal research, the main criterion for selecting the number of persons and measurement occasions is power (see, e.g., Lafit et al. 2021). Power analysis focuses on comparing two nested models (i.e., the unrestricted model and the restricted model that corresponds to the null hypothesis), that usually differ with respect to one model parameter only. This criterion might be restrictive when researchers are interested in comparing a set of competing non-nested models, or are interested in the development of predictive models. We therefore argue that predictive accuracy can also be used to inform sample size planning, where researchers can assess how many observations are needed to have low estimation bias and variance and thus good prediction. The latter can be done in an absolute (i.e., prediction accuracy is better than a set threshold) or relative way (i.e., better prediction than a competing, incorrect model). Further elucidating such similarities and differences between power analysis and predictive accuracy would be useful.

7.3. Conclusion

In conclusion, in this paper, we conducted three simulation studies to grasp the effect of different data characteristics that show up in psychological applications, on assessing predictive accuracy and quantifying how well person-specific and multilevel VAR(1) models generalize to unseen data. Our analyses suggest that pooling information across individuals and using multilevel techniques prevents overfitting, and outperforms person-specific VAR(1) models, even when the data include clusters of individuals with similar dynamics. Moreover, the results of our simulations demonstrated that when contemporaneous multicollinearity is likely to be present in the data, multilevel VAR models yield good predictive performance. However, further research is needed to evaluate the extent to which dimension reduction techniques can be used to characterize psychological dynamics reliably. Finally, we identified pressing questions regarding study design and model selection that deserve future attention.

Acknowledgments

The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

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 (C14/19/054; iBOF/21/090) awarded to E. Ceulemans.

Appendix

Performance Measures to Evaluate Estimation Accuracy. While we cannot directly estimate the mean squared bias or variance of the predictions in our CV setting, an easy way to shed some light on the relation between predictive accuracy and estimation accuracy is to investigate how well person-specific and multilevel VAR(1) models can accurately estimate the elements of the true transition matrix $\mathbf{\Psi}_i$, underlying our simulated data. Specifically, using the complete data sets (i.e., without further splitting them in a training and test part), we can compute the mean squared estimation error for the autoregressive effects as follows:

$$\mathrm{MSE}_{\psi_{jj}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{P}\sum_{j=1}^{P}\left(\hat{\psi}_{ijj} - \psi_{ijj}\right)^2\right) \tag{12}$$

where ψ_{ijj} and $\hat{\psi}_{ijj}$ denote the true and estimated autoregressive effects. Additionally, we can compute the mean squared estimation errors for the cross-regressive effects:

$$\mathrm{MSE}_{\psi_{jk}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{P(P-1)}\sum_{j=1}^{P}\sum_{\substack{k=1\\k\neq j}}^{P}\left(\hat{\psi}_{ijk} - \psi_{ijk}\right)^2\right) \tag{13}$$

where ψ_{ijk} and $\hat{\psi}_{ijk}$ denote the true and estimated cross-regressive effects. We expect that these two MSE measures are strongly correlated to how well a considered modeling approach estimates the target function $f(\mathbf{Y}_{it-1})$. We will do that for Study I and report the results below (Tables 9 and 10). Figures 5, 6, 7, 8, 9, 10, 11, and 12 display the histograms of the distribution of the auto- and cross-regressive effects across persons when data are generated from a multilevel VAR(1) model with random effects, a person-specific VAR(1) model, or a cluster-specific VAR(1) model.

TABLE 9.

Simulation results of study I. The mean squared estimation error (standard deviation in parentheses) for the autoregressive effects when data are generated using person-specific and multilevel VAR(1) models.

Population model	Method	Number of variables			Number of persons			Measurement occasions		
		$P = 2$	$P = 4$	$P = 8$	$N = 20$	$N = 60$	$N = 120$	$T = 50$	$T = 100$	$T = 200$
VAR	AR	0.001	0.019	0.201	0.072	0.074	0.074	0.042	0.075	0.103
		(0.001)	(0.007)	(0.070)	(0.098)	(0.100)	(0.100)	(0.052)	(0.093)	(0.127)
	MAR.FE	0.003	0.038	0.279	0.103	0.108	0.109	0.081	0.109	0.129
		(0.001)	(0.007)	(0.058)	(0.126)	(0.127)	(0.129)	(0.092)	(0.127)	(0.152)
	MAR.RE	0.002	0.027	0.236	0.086	0.089	0.090	0.063	0.090	0.112
		(0.001)	(0.005)	(0.060)	(0.109)	(0.111)	(0.112)	(0.073)	(0.107)	(0.136)
	MVAR.FE	0.000	0.001	0.005	0.002	0.002	0.002	0.001	0.002	0.004
		(0.001)	(0.000)	(0.004)	(0.003)	(0.003)	(0.003)	(0.001)	(0.002)	(0.004)
	MVAR.RE	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
		(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
	VAR	0.001	0.002	0.003	0.003	0.002	0.002	0.005	0.001	0.000
		(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.000)
MVAR.FE	AR	0.001	0.018	0.232	0.083	0.083	0.083	0.049	0.085	0.115
		(0.001)	(0.006)	(0.077)	(0.115)	(0.114)	(0.114)	(0.063)	(0.108)	(0.146)
	MAR.FE	0.002	0.026	0.306	0.110	0.112	0.112	0.085	0.114	0.135
		(0.001)	(0.004)	(0.062)	(0.142)	(0.143)	(0.143)	(0.104)	(0.142)	(0.170)
	MAR.RE	0.002	0.026	0.271	0.099	0.100	0.100	0.072	0.101	0.125
		(0.001)	(0.005)	(0.065)	(0.127)	(0.127)	(0.127)	(0.087)	(0.124)	(0.156)
	MVAR.FE	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
		(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
	MVAR.RE	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
	VAR	0.001	0.002	0.003	0.002	0.002	0.002	0.005	0.001	0.000
		(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.000)

TABLE 9.
continued

Population model	Method	Number of variables			Number of persons			Measurement occasions		
		$P = 2$	$P = 4$	$P = 8$	$N = 20$	$N = 60$	$N = 120$	$T = 50$	$T = 100$	$T = 200$
MVAR.RE	AR	0.002	0.019	0.074	0.033	0.031	0.031	0.021	0.033	0.041
		(0.002)	(0.008)	(0.022)	(0.035)	(0.033)	(0.033)	(0.022)	(0.033)	(0.040)
	MAR.FE	0.009	0.064	0.197	0.088	0.090	0.092	0.072	0.092	0.107
		(0.007)	(0.023)	(0.048)	(0.085)	(0.085)	(0.085)	(0.066)	(0.085)	(0.098)
	MAR.RE	0.003	0.028	0.094	0.042	0.041	0.041	0.037	0.042	0.046
		(0.003)	(0.008)	(0.017)	(0.041)	(0.039)	(0.039)	(0.034)	(0.040)	(0.044)
	MVAR.FE	0.003	0.013	0.069	0.024	0.029	0.032	0.020	0.029	0.036
		(0.003)	(0.007)	(0.027)	(0.027)	(0.034)	(0.037)	(0.022)	(0.032)	(0.041)
	MVAR.RE	0.001	0.001	0.001	0.002	0.001	0.000	0.001	0.001	0.001
		(0.001)	(0.001)	(0.001)	(0.002)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
	VAR	0.002	0.003	0.004	0.004	0.002	0.002	0.005	0.002	0.001
		(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)	(0.001)

TABLE 10.

Simulation results of study I. The mean squared estimation error (standard deviation in parentheses) for the cross-regressive effects when data are generated using person-specific and multilevel VAR(1) models.

Population model	Method	Number of variables			Number of persons			Measurement occasions		
		$P = 2$	$P = 4$	$P = 8$	$N = 20$	$N = 60$	$N = 120$	$T = 50$	$T = 100$	$T = 200$
VAR	MVAR.FE	0.033	0.018	0.007	0.019	0.022	0.018	0.020	0.019	0.019
		(0.006)	(0.002)	(0.001)	(0.010)	(0.014)	(0.009)	(0.012)	(0.011)	(0.011)
	MVAR.RE	0.033	0.018	0.006	0.019	0.021	0.017	0.019	0.019	0.019
		(0.006)	(0.002)	(0.001)	(0.010)	(0.014)	(0.010)	(0.012)	(0.012)	(0.011)
	VAR	0.034	0.018	0.006	0.019	0.021	0.018	0.020	0.019	0.019
		(0.006)	(0.002)	(0.001)	(0.011)	(0.014)	(0.010)	(0.012)	(0.012)	(0.011)
MVAR.FE	MVAR.FE	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
	MVAR.RE	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
	VAR	0.028	0.012	0.006	0.016	0.015	0.015	0.016	0.015	0.015
		(0.004)	(0.001)	(0.001)	(0.010)	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)
MVAR.RE	MVAR.FE	0.001	0.002	0.004	0.003	0.002	0.002	0.002	0.002	0.002
		(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	MVAR.RE	0.001	0.001	0.001	0.002	0.001	0.000	0.001	0.001	0.001
		(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
	VAR	0.028	0.013	0.007	0.017	0.016	0.016	0.017	0.016	0.016
		(0.007)	(0.002)	(0.001)	(0.011)	(0.009)	(0.009)	(0.010)	(0.010)	(0.009)

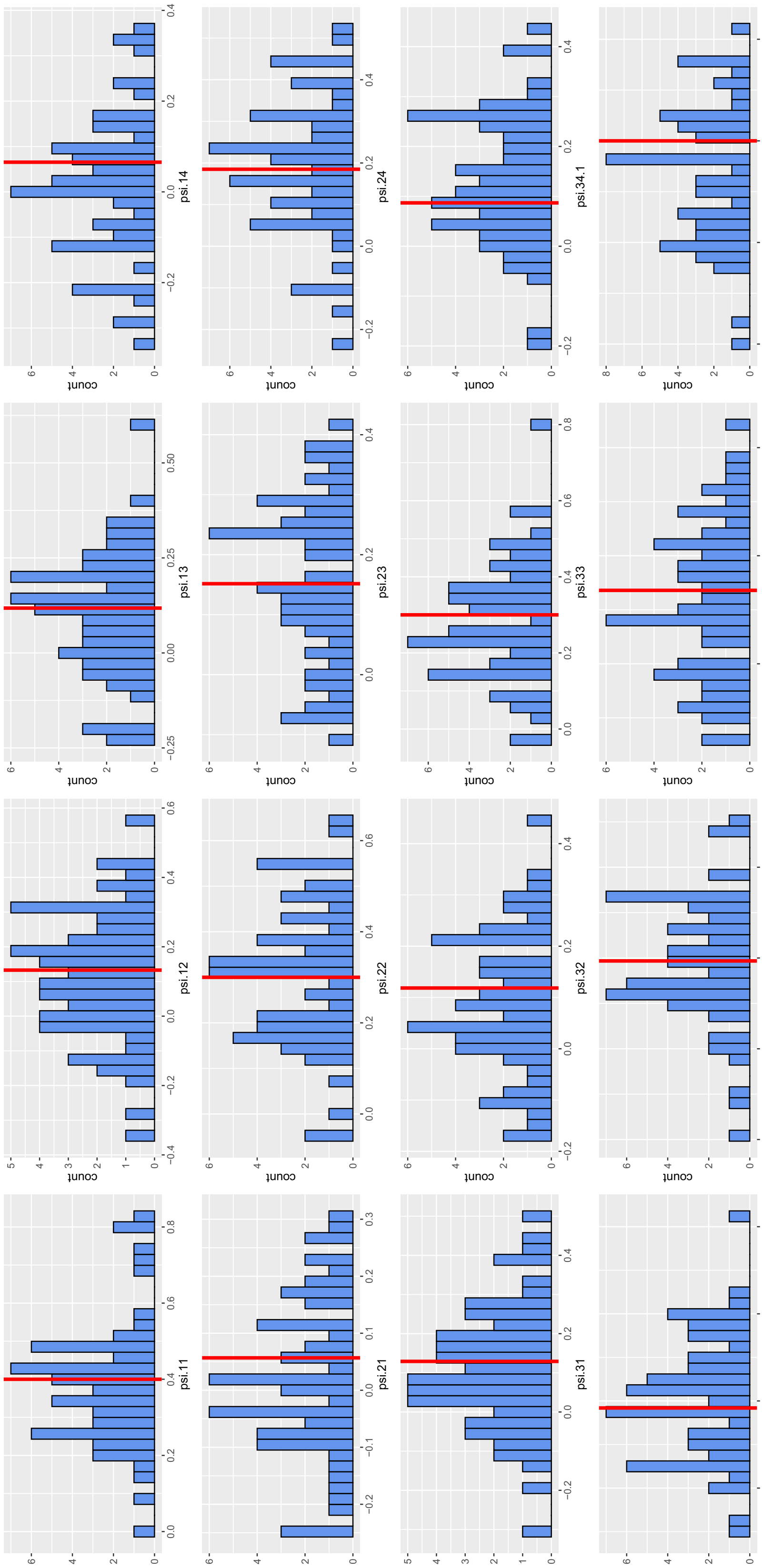

FIGURE 5.
Distribution of the elements of the transition matrix for a multilevel VAR(1) model with random effects with 4 variables and 60 individuals. The vertical lines represent the fixed effects.

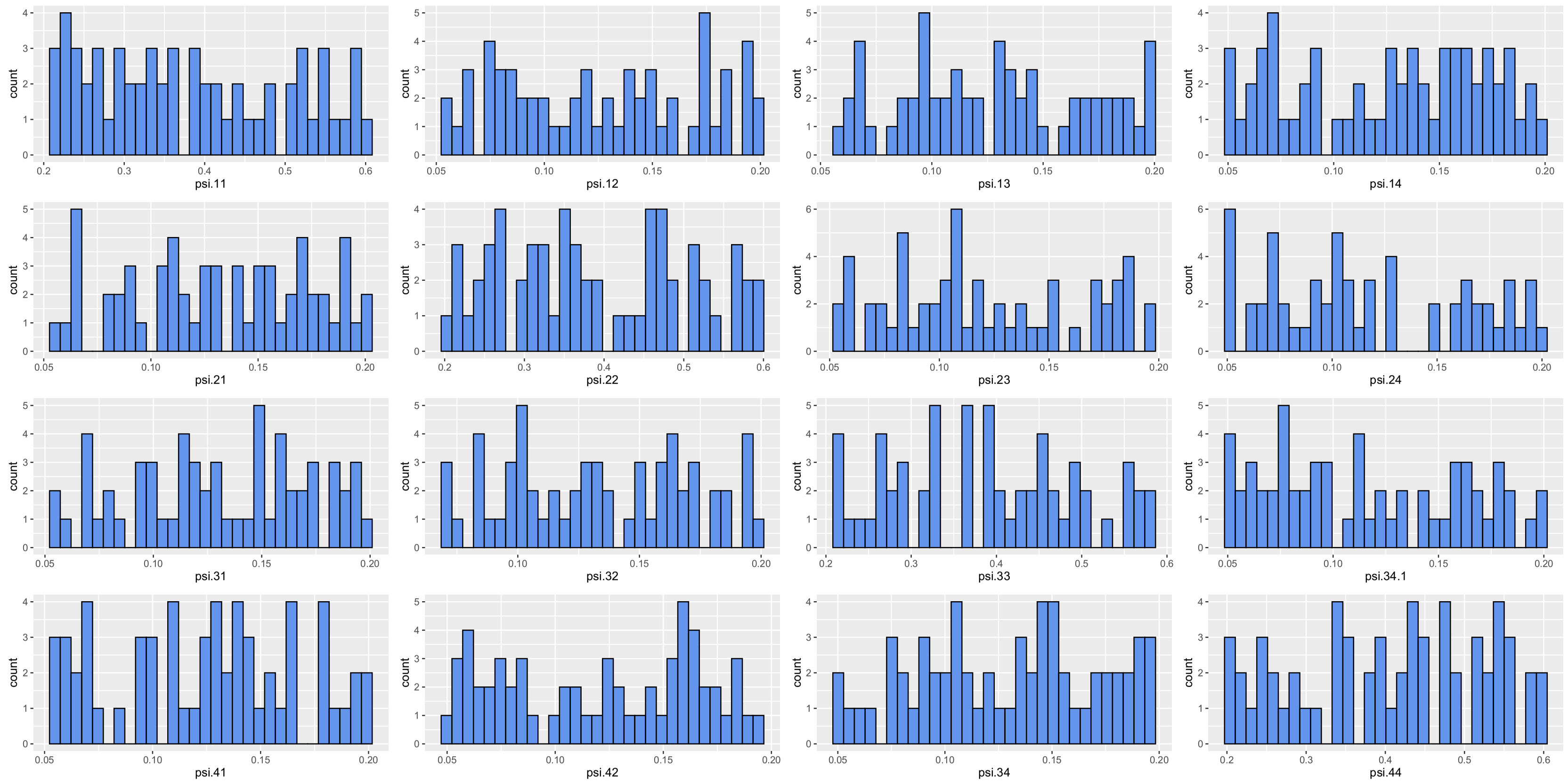

FIGURE 6.
Distribution of the elements of the transition matrix for a person-specific VAR(1) model with 4 variables and 60 individuals.

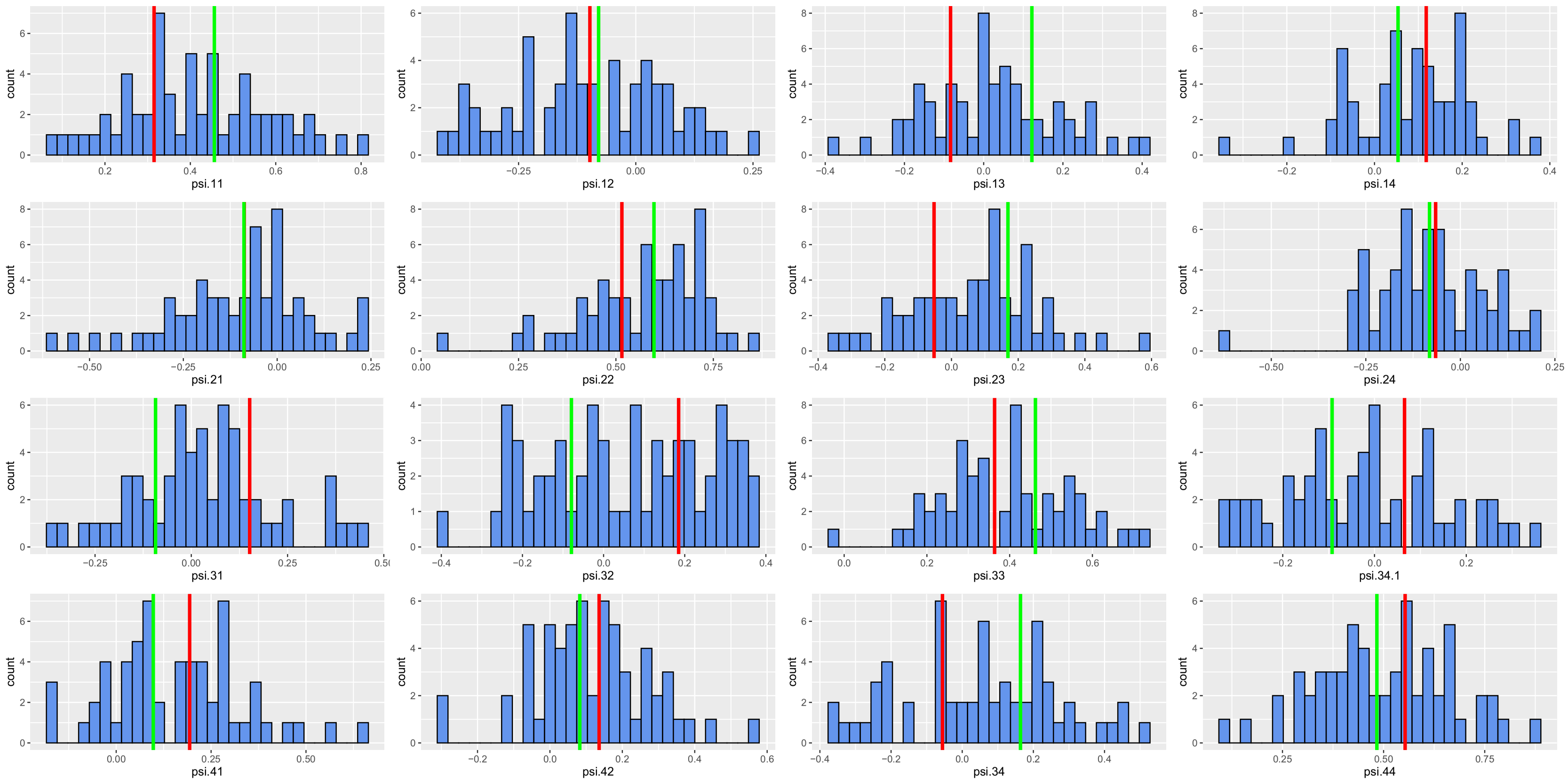

FIGURE 7.
Distribution of the elements of the transition matrix for a cluster-specific VAR(1) model with random effects with 4 variables, 2 clusters of equal size, and 60 individuals. The vertical lines represent the fixed effects for each cluster.

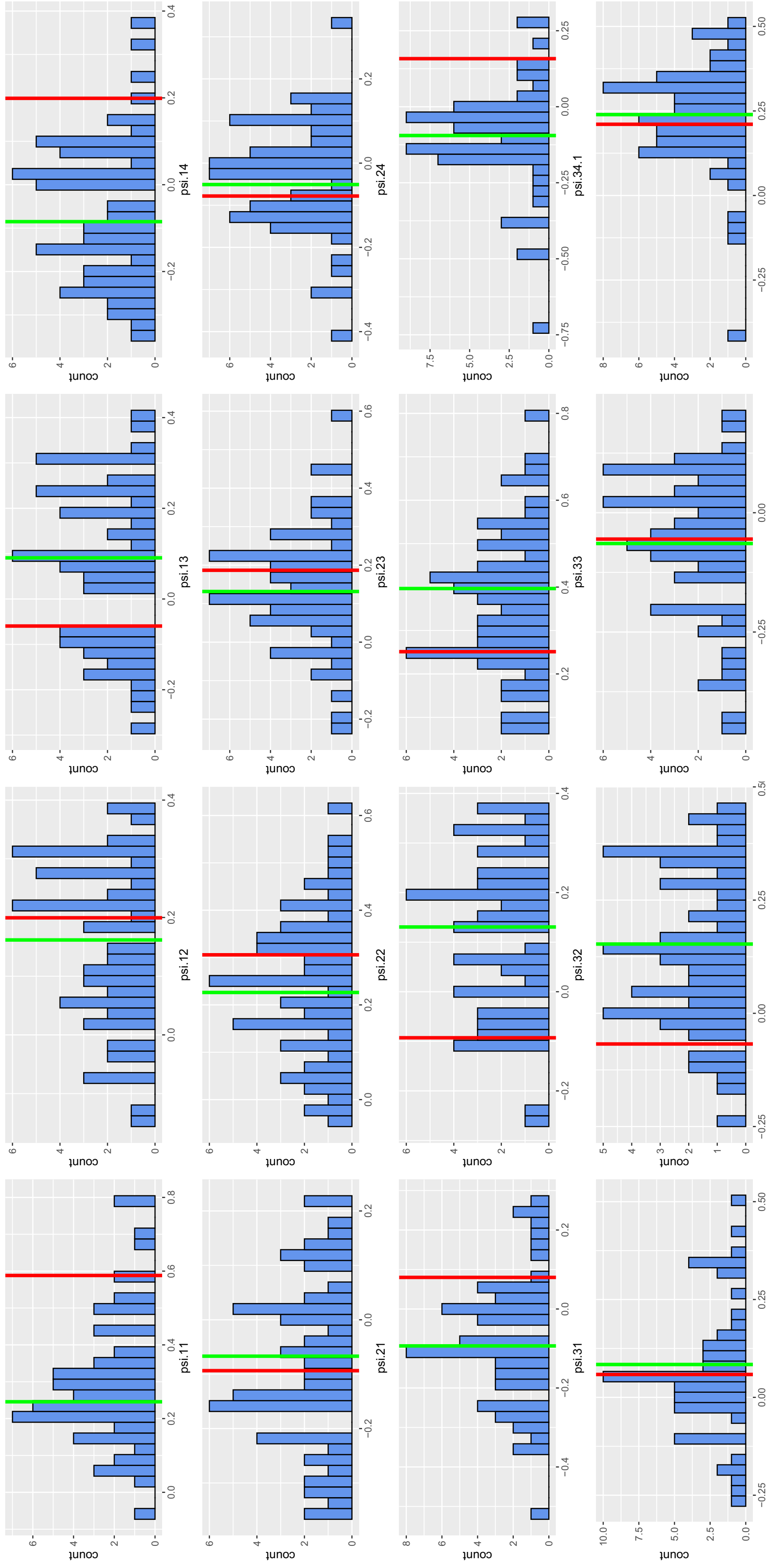

FIGURE 8.
Distribution of the elements of the transition matrix for a cluster-specific VAR(1) model with random effects with 4 variables, 2 clusters with one cluster including 10% of the persons, and 60 individuals. The vertical lines represent the fixed effects for each cluster.

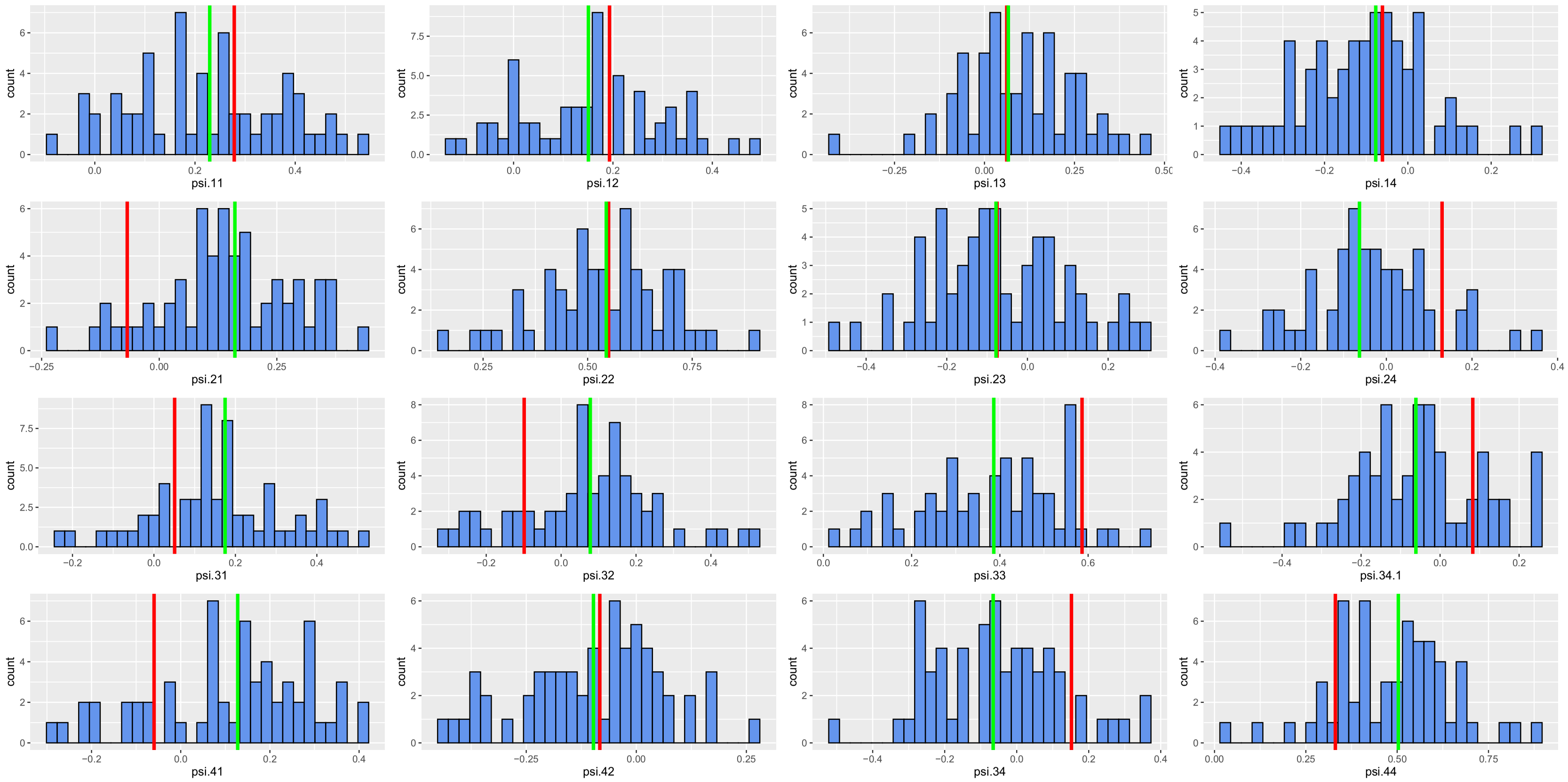

FIGURE 9.
Distribution of the elements of the transition matrix for a cluster-specific VAR(1) model with random effects with 4 variables, 2 clusters with one cluster including 60% of the persons, and 60 individuals. The vertical lines represent the fixed effects for each cluster.

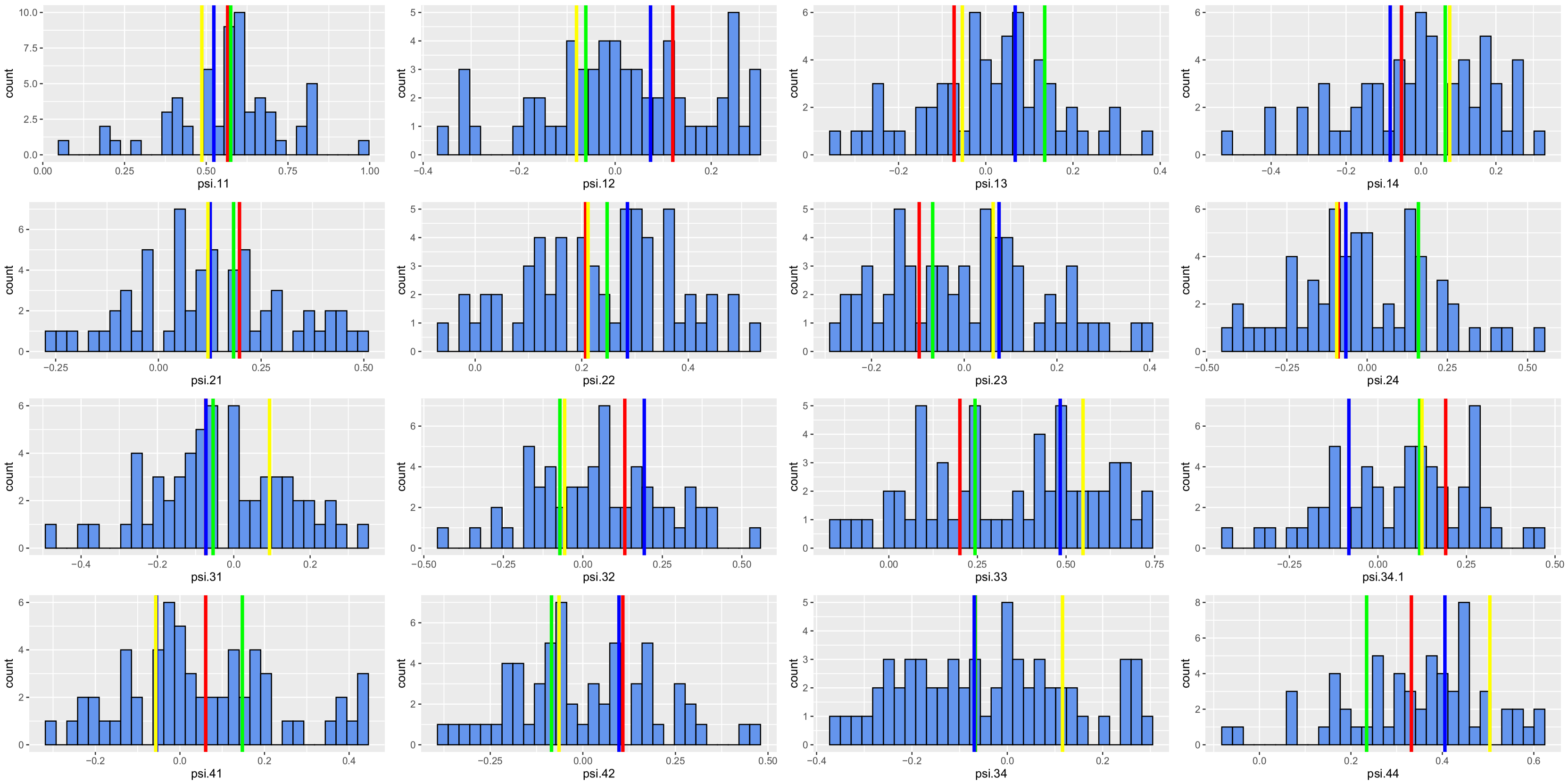

FIGURE 10.
Distribution of the elements of the transition matrix for a cluster-specific VAR(1) model with random effects with 4 variables, 4 clusters of equal size, and 60 individuals. The vertical lines represent the fixed effects for each cluster.

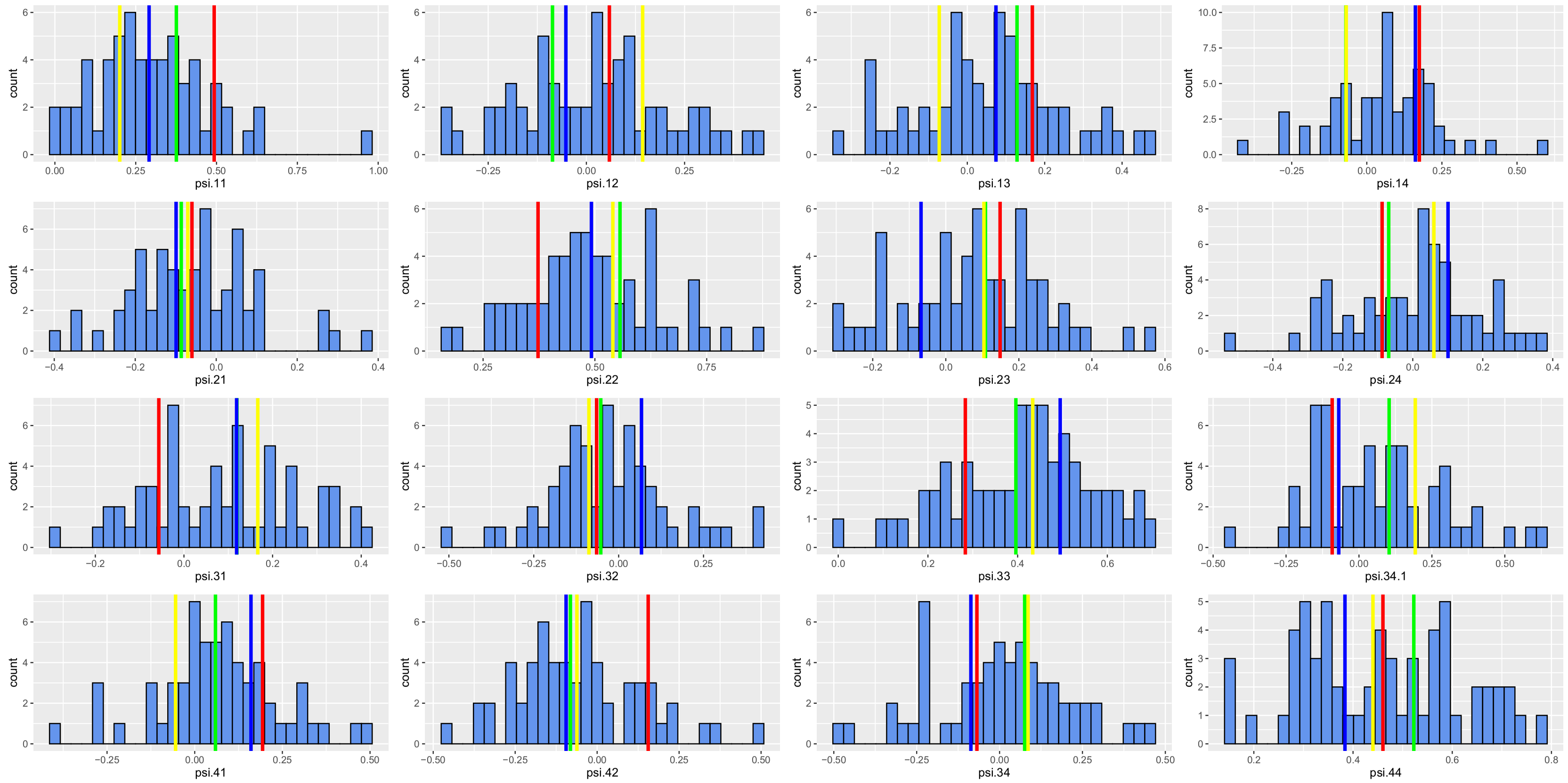

FIGURE 11.
Distribution of the elements of the transition matrix for a cluster-specific VAR(1) model with random effects with 4 variables, 4 clusters with one cluster including 10% of the persons, and 60 individuals. The vertical lines represent the fixed effects for each cluster.

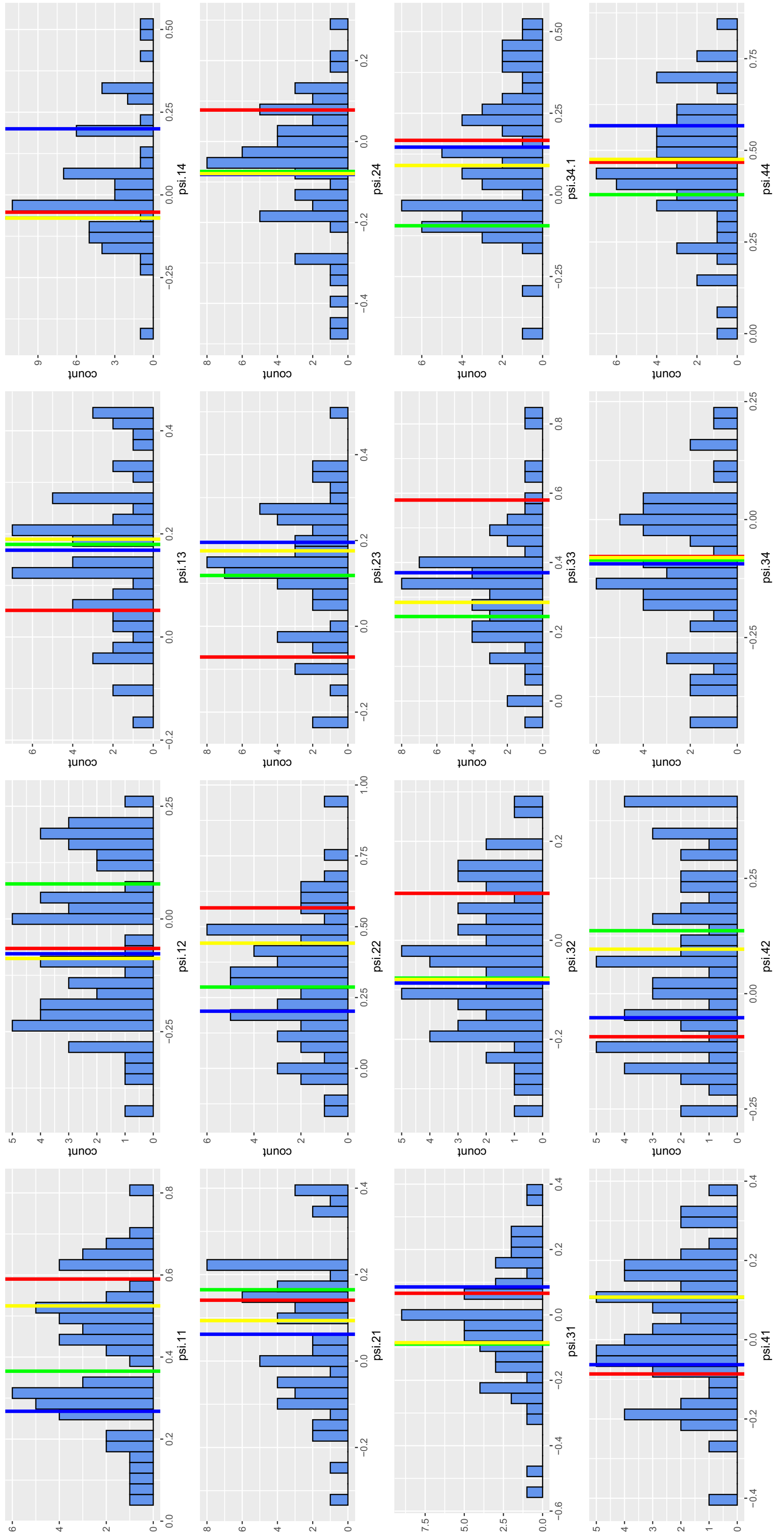

FIGURE 12.
Distribution of the elements of the transition matrix for a cluster-specific VAR(1) model with random effects with 4 variables, 4 clusters with one cluster including 60% of the persons, and 60 individuals. The vertical lines represent the fixed effects for each cluster.

References

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

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.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Kelman, T., Simon, A. B., Noack, A., Hatherly, M., & Bouchet-Valat, M. (2016). Dmbates/Mixedmodels.Jl: Drop Julia V0.4.X and earlier support. *Zenodo*.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015a). *Parsimonious mixed models*. arXiv preprint arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*(1), 68–98.

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

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., & 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., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8*(4), e60188.

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between-person and within-person structures of affect are a matter of degree. *European Journal of Personality, 29*(1), 55–71.

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of autoregressive moving average time series models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 415–452). Mahwah, NJ: Lawrence Erlbaum Associates.

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018a). VAR (1) based models do not always outpredict AR (1) models in typical psychological applications. *Psychological Methods, 23*(4), 740–756.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016a). Clustering vector autoregressive models: Capturing qualitative differences in within-person dynamics. *Frontiers in Psychology, 7,* 1540.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016b). Using raw VAR regression coefficients to build networks can be misleading. *Multivariate Behavioral Research, 51*(2–3), 330–344.

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

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies, 21*(4), 1509–1531.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Ceulemans, E., & Kiers, H. A. (2009). Discriminating between strong and weak structures in three-mode principal component analysis. *British Journal of Mathematical and Statistical Psychology, 62*(3), 601–620.

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology, 59*(1), 133–150.

Ceulemans, E., Timmerman, M. E., & Kiers, H. A. (2011). The CHull procedure for selecting among multilevel component solutions. *Chemometrics and Intelligent Laboratory Systems, 106*(1), 12–20.

Ceulemans, E., & Van Mechelen, I. (2005). Hierarchical classes models for three-way three-mode binary data: Interrelations and model selection. *Psychometrika, 70*(3), 461–480.

Ceulemans, E., Wilderjans, T. F., Kiers, H. A. L., & Timmerman, M. E. (2016). MultiLevel simultaneous component analysis: A computational shortcut and software package. *Behavior Research Methods, 48,* 1008–1020.

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects. *Political Science Research and Methods, 3*(2), 399–408.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Milton Park: Routledge.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement, 70*(6), 885–901.

Eisele, G., Lafit, G., Vachon, H., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). *Affective structure, measurement invariance, and reliability across different experience sampling protocols*.

Ernst, A. F., Timmerman, M. E., Jeronimus, B. F., & Albers, C. J. (2019). Insight into individual differences in emotion dynamics with clustering. *Assessment*, first online.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics (Vol. 1(10)). New York: Springer.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery, 1*(1), 55–77.

Gates, K. M., & Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*(1), 310–319.

Gelman, A. (2005). Analysis of variance-why it is more important than ever. *Annals of Statistics, 33*(1), 1–53.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Hoboken: Wiley.

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

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

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.

Kiers, H. A., & Smilde, A. K. (2007). A comparison of various methods for multivariate regression with highly collinear variables. *Statistical Methods and Applications, 16*(2), 193–228.

Kiers, H. A. L., & ten Berge, J. M. F. (1994a). Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *British Journal of Mathematical and Statistical Psychology, 47,* 109–126.

Kiers, H. A. L., & ten Berge, J. M. F. (1994b). The Harris-Kaiser independent cluster rotation as a method for rotation to simple component weights. *Psychometrika, 59,* 81–90.

Krone, T., Albers, C. J., Kuppens, P., & Timmerman, M. E. (2018). A multivariate statistical model for emotion dynamics. *Emotion, 18*(5), 739–754.

Krone, T., Albers, C. J., & Timmerman, M. E. (2016). Comparison of estimation procedures for multilevel AR (1) models. *Frontiers in Psychology, 7,* 486.

Krone, T., Albers, C. J., & Timmerman, M. E. (2017). A comparative simulation study of AR(1) estimators in short time series. *Quality & Quantity, 51*(1), 1–21.

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,* 380.

Lafit, G., Adolf, J., 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. In *Advances in methods and practices in psychological science*.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *New directions for methodology of social and behavioral science* (pp. 41–56). San Francisco: Jossey-Bass.

Liu, S. (2017). Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels. *British Journal of Mathematical and Statistical Psychology, 70*(3), 480–498.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340–364.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer.

Mansueto, A. C., Wiers, R., van Weert, J. C., Schouten, B. C., & Epskamp, S. (2020). *Investigating the feasibility of idiographic network models*.

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.

Merz, E. L., & Roesch, S. C. (2011). Modeling trait and state variation using multilevel factor analysis with PANAS daily diary data. *Journal of Research in Personality, 45*(1), 2–9.

Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201–218.

Morren, M., Van Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: a systematic review. *European Journal of Pain, 13*(4), 354–365.

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science, 28*(2), 135–167.

Muthén, B., & Muthén, B. O. (2009). *Statistical analysis with latent variables*. New York, NY: Wiley.

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry, 17*(2), 123–132.

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. *Journal of Medical Internet Research, 21*(2), e11398.

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

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Schepers, J., Ceulemans, E., & Van Mechelen, I. (2008). Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. *Journal of Classification, 25*(1), 67.

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 495–515.

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

Sels, L., Ceulemans, E., Bulteel, K., & Kuppens, P. (2016). Emotional interdependence and well-being in close relationships. *Frontiers in Psychology, 7,* 283.

Song, H., & Zhang, Z. (2014). Analyzing multiple multivariate time series data using multilevel dynamic factor models. *Multivariate Behavioral Research, 49*(1), 67–77.

Timmerman, M. E., & Kiers, H. A. L. (2003). Four simultaneous component models of multivariate time series for more than one subject to model intraindividual and interindividual differences. *Psychometrika, 86,* 105–122.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9,* 151–176.

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research, 21*(12), e14475.

Wainer, H. (1976). Estimating coefficients in linear models: It dont make no nevermind. *Psychological Bulletin, 83*(2), 213.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine, 44*(7), 1349–1360.

Wigman, J. T. W., Van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., & 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. *Psychological Medicine, 45*(11), 2375–2387.

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull: A generic convex hull based model selection method. *Behavior Research Methods, 45*(1), 1–15.

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

Zautra, A. J., Affleck, G. G., Tennen, H., Reich, J. W., & Davis, M. C. (2005). Dynamic approaches to emotions and stress in everyday life: Bolger and Zuckerman reloaded with positive as well as negative affects. *Journal of Personality, 73*(6), 1511–1538.

Manuscript Received: 13 JUL 2020
Final Version Received: 13 JUL 2021
Accepted: 2 AUG 2021
Published Online Date: 1 NOV 2021