Clustering Individuals on Limited Features of a Vector Autoregressive Model

Article *in* Multivariate Behavioral Research · May 2020

DOI: 10.1080/00273171.2020.1767532

4 authors:

Keisuke Takano
Ludwig-Maximilians-University of Munich
100 PUBLICATIONS 1,142 CITATIONS
SEE PROFILE

Mina Stefanovic
Ludwig-Maximilians-University of Munich
2 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Tabea Rosenkranz
Ludwig-Maximilians-University of Munich
11 PUBLICATIONS 54 CITATIONS
SEE PROFILE

Thomas Ehring
Ludwig-Maximilians-University of Munich
150 PUBLICATIONS 6,197 CITATIONS
SEE PROFILE

Clustering Individuals on Limited Features of a Vector Autoregressive Model

Keisuke Takano

Mina Stefanovic

Tabea Rosenkranz

Thomas Ehring

Department of Psychology, LMU Munich, Germany

Correspondence concerning this article should be addressed to Keisuke Takano. Department of Psychology, Division of Clinical Psychology and Psychotherapy, Ludwig-Maximilians-University (LMU) Munich, Leopoldstrasse.13, 80802 Munich, Germany. Telephone: +49 89 2180 5177. Fax: +49 89 2180 5224. E-mail: Keisuke.Takano@psy.lmu.de

This is an accepted version of the manuscript. To cite: Takano K, Stefanovic M, Rosenkranz T, Ehring T. Clustering Individuals on Limited Features of a Vector Autoregressive Model [published online ahead of print, 2020 May 20]. Multivariate Behav Res. 2020;1 - 19. doi:10.1080/00273171.2020.1767532

Abstract

Dynamical interplays in emotions have been investigated using vector autoregressive (VAR) models, whose estimates can be used to cluster participants into unknown groups. The present study evaluated a clustering algorithm, the alternating least square (ALS) algorithm, for accuracy in predicting individual group membership. We systematically manipulated (a) the number of variables in a model, (b) the size of group differences in regression coefficients, and (c) the number of regression coefficients that vary across the groups (i.e., effective features). The ALS algorithm works reliably when there are at least 5 effective features with very large group differences in a 5-variable model; and 9 effective features with very large group differences in a 10-variable model. These findings suggest that the ALS algorithm is sensitive to group differences that are present only in several coefficients of a VAR model, but that the group differences have to be large. We also found that the ALS algorithm outperforms another clustering method, Gaussian mixture modeling. The ALS algorithm was further evaluated with unbalanced sample sizes between groups and with a greater number of groups in data (Study 2). A real data application was provided to illustrate how to interpret the detected group differences (Study 3).

Keywords: vector autoregressive model, emotion dynamics, clustering, experience sampling method

Clustering Individuals on Limited Features of a Vector Autoregressive Model

Dynamic interplays between symptoms and/or emotions have been gaining attention in recent years. Symptom connectivity is expected to represent or even predict individual courses of development in psychopathology and to reveal how symptom clusters emerge within and across different psychopathologies (e.g., Borsboom, 2017; Wichers, Schreuder, Goekoop, & Groen, 2019). Studies investigating dynamic processes between different emotions in daily life (e.g., Pe & Kuppens, 2012) showed that a current emotion influences the experience of another emotion at the next time point even after controlling for current levels of that emotion. Such short-term prospective associations of symptoms and emotions have been studied in various types of psychopathology (Wigman et al., 2015), such as depression (Pe et al., 2015), post-traumatic stress disorder (Greene, Gelkopf, Epskamp, & Fried, 2018), and schizophrenia (Klippel et al., 2018). These findings hold implications regarding how psychopathology is characterized by the chains and sequences (or even causalities) of symptoms and/or emotions.

Vector Autoregressive (VAR) Model

A Vector Autoregressive (VAR) Model has been used to study person-specific dynamic interplays between psychopathological symptoms or emotions, which is typically fitted on intensive longitudinal data acquired through the experience sampling method (ESM) or ecological momentary assessment. In a typical ESM study, each participant receives around 5–10 signals per day on their smart device and is asked to report their current mood, thoughts, and behavior in response. Such ESM data are a subject of time series analyses, including autoregressive (AR) models to capture the persistency of affective states (e.g., Jahng, Wood, & Trull, 2008) and VAR models to estimate prospective associations between multiple time series. A VAR model is formulated as a set of regressions where variables are predicted by past states of the variables (including the lagged outcome variable). In the psychopathology literature, it is common to use a lag-1 model, i.e., VAR(1), which can be described as follows (e.g., Bulteel, Tuerlinckx, Brose, & Ceulemans, 2016; Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016):

$$\boldsymbol{Y}_{ti} = \boldsymbol{c}_i + \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-1i} + \boldsymbol{\omega}_{ti} \qquad (1)$$

where $\boldsymbol{Y}_{ti}$ is an M-length vector containing the observed values of M-number variables at time t for participant i; $\boldsymbol{c}_i$ is also a vector, representing the intercepts of the M variables; $\boldsymbol{\Phi}_i$ is an $M \times M$ matrix containing the regression coefficients of the lagged variables $\boldsymbol{Y}_{t\text{-}1i}$; and $\boldsymbol{\omega}_{it}$ is an M-length vector of innovations (errors) at time t, which are independent across time points. Note that the number of observations, T_i, can vary across individuals. The entries of the diagonal of $\boldsymbol{\Phi}_i$ are the autoregression coefficients reflecting resistance to change (or emotional inertia in the case of studies investigating emotional dynamics); these coefficients are known to, e.g., predict current and future levels of depressive symptoms (Kuppens, Allen, & Sheeber, 2010; Kuppens et al., 2012). A cross-regression coefficient represents the prospective effect of one variable on another (e.g., Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015)[1].

As ESM data often have a nested structure due to repeated measurements, individual VAR models have been specified by individualistic (or single-subject) estimation or hierarchical modeling; therefore, the auto- and cross-regression coefficients are typically assumed to vary across individuals. For the individualistic estimation, a VAR model can be fitted on each participant, which yields individual regression coefficients (e.g., Zheng, Wiebe, Cleveland, Molenaar, & Harris, 2013). In general, this individualistic approach is straightforward and easy to

[1] A VAR model can be mapped onto a network diagram with nodes representing variables and with edges representing connectivity (cross-regression coefficients) between the variables. In ESM data, a "network" can be defined for each individual, which provides indices that characterize individual psychological networks (e.g., centrality) in analogy to a social network. There is, however, an ongoing debate on how to conceptualize and configure an individualistic psychological network, for which the direct application of the social-network measures may not be appropriate due to the conceptual differences (Bringmann et al., 2019). Although such a psychological network is out of our focus, the network analysis (on VAR estimates) can be an interesting direction to model the interrelations of symptoms and/or emotions.

implement, but sometimes yields unreliable estimates when the number of data points is limited for a person (e.g., Katahira, 2016). This is mainly because the individualistic estimation does not consider the group-level population distribution of model parameters, which can be an extra source of information that improves estimation precision at the individual level (e.g., Ahn, Krawitz, Kim, Busmeyer, & Brown, 2011).

The hierarchical approach explicitly models the group-level population distribution in the frequentist (Bringmann et al., 2013, 2016) or hierarchical Bayes framework (Asparouhov et al., 2018; Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011). In ESM/VAR contexts, the frequentist approach has been used more frequently. Here, a VAR structure is specified by estimating univariate hierarchical models for every variable. In each univariate model, regression coefficients have random effects, which allow the auto- and cross-regression coefficients to vary across individuals (See Study 3 for more details). These estimation techniques are also summarized in Epskamp et al. (2018), and are implemented as a R package, mlVAR (Epskamp, Deserno, & Bringmann, 2019).

Clustering Individuals on the Basis of a VAR Structure

Given that there are potentially meaningful individual differences in the VAR structure, it is clinically relevant to explore whether groups of individuals can be identified who share (maladaptive) features of emotion-to-emotion (or symptom-to-symptom) interrelationships; for example, there could be a group of individuals who have a strong association between momentary rumination/worry and negative affect that are assessed via ESM. This strong connectivity is regarded as a precursor or an early warning sign that predicts a near-future onset of depression (Wichers, Groot, & Group, 2016; Wichers et al., 2019). Recent methodological studies have proposed algorithms to cluster participants into unknown groups on the basis of the similarity of person-specific VAR (and other type of regression) estimates. This clustering approach is expected (a) to identify a group of people who share the same or similar characteristics in the associations between given symptoms and emotions; and (b) to relate the group labels to risks of future psychopathology and prognosis of psychological treatment.

Although our focus was specifically on clustering individuals on the basis of VAR estimates, there have been a number of approaches to cluster time series in general. Under the taxonomy of Liao (2005), the time-series clustering can be categorized into three different approaches, which are based on (a) the raw data, (b) features extracted from the raw time series, and (c) model parameters such as regression coefficients and residuals. Because we were most interested in the dynamic interplays between multiple time series, the current study had a particular focus on (c), i.e., clustering based on model parameters, which are the auto- and cross-regression coefficients in our case. VAR-based clustering can be seen as a multivariate extension of autoregressive (AR) based clustering approaches. Typically, AR-based clustering targets autoregressive representations of a univariate time series, which are used as features for a specific clustering technique such as finite mixture modeling (e.g., D'Urso, Di Lallo, & Maharaj, 2013; Frühwirth-Schnatter & Kaufmann, 2008).

The mixture-modeling approach assumes that regression coefficients are drawn from a mixture of multiple distributions; e.g., the finite Gaussian mixture model (GMM) assumes that each group of participants follows different normal distributions, and can estimate probabilistic group membership for each participant. Although this type of clustering has been implemented in various software packages (R package flexmix, which is for non-hierarchical data; Leisch, 2019), it typically (as noted above) targets univariate models such as standard or multilevel regressions, including growth curve models (e.g., proc traj; Jones, Nagin, & Roeder, 2001). Extending this GMM approach into multi time series is computationally demanding (if not impossible), because the number of regression coefficients increases exponentially as the number of variables in a model increases; furthermore, each regression coefficient may follow different group distributions, which require additional free parameters to estimate.

A possible workaround to apply GMM in the individual (or hierarchical) VAR context is the use of a two-step estimation scheme: i.e., (a) to fit a generative (or VAR) model to each participant, and (b) to submit the individual estimates of the model parameters to any clustering analysis. An example implementation can be found in Zheng et al. (2013), where the VAR(1) was fitted to each individual to model the interplays between substance use craving and negative affect and tabacco use. The estimated person-specific regression coefficients were then used for hierarchical clustering. Brodersen et al. (2014) used GMM to cluster individuals on the person-specific estimates of neural dynamics between brain regions of interest. They first processed the brain-imaging time series by a generative model (here: dynamic causal model) to obtain person-specific parameter estimates that represent neural connectivity. These parameter estimates were used as a similarity metric for further GMM clustering. A more recent study (Ernst, Timmerman, Jeronimus, & Albers, 2019) combined the individualistic estimation of VAR and clustering by GMM, suggesting that GMM has clear advantages over the traditional hierarchical clustering and/or non-probabilistic clustering methods[2]; for example, GMM can directly model variation within a cluster, allowing for flexible assumptions on the orientation and distribution of different clusters. Also, information criteria are available to determine the number of clusters. Possible concerns were that (a) clustering performance could be affected by precision of the individual estimates of VAR parameters and that (b) GMM might need further dimension reduction or

[2] Performance of the hierarchical clustering on individual VAR estimates was already examined by Bulteel et al. (2016). The outputs of the hierarchical clustering were used as initial values for the ALS optimization, which are known to be less accurate than the final predictions of the ALS algorithm. Therefore, we did not include this approach in the current evaluation.

feature selection because the feature space for VAR clustering (i.e., **Φ** matrix) likely has a large number of random dimensions that do not contribute to the clustering (Scrucca & Raftery, 2014).

An alternative clustering approach for VAR models is the alternating least squares (ALS) algorithm, proposed by Bulteel et al. (2016). This algorithm does not assume a specific distribution for a regression coefficient; instead, it searches for the best partitioning of participants by minimizing prediction errors (or residuals) of VAR models that are separately fitted on the given groups. The algorithm mainly consists of three steps of optimization. First, group membership is tentatively assigned to each individual. This initial membership is given by a certain criterion on the basis of other clustering methods such as hierarchical clustering. Second, a VAR model is estimated for each group but not on each individual. This means that data from the individuals in the same group is combined to estimate a group-wise VAR model; therefore, $\boldsymbol{c}_i$ and $\boldsymbol{\Phi}_i$ only vary across groups but are identical within a group. The models' prediction errors (residuals) are evaluated to update individual group membership. Here the loss function is defined as a sum of prediction errors (i.e., the differences between the observed vs. predicted *y* values) across time points and participants:

$$L_k = \sum_{i}^{I} \sum_{t}^{T_i} (\boldsymbol{y}_{it} - \widehat{\boldsymbol{y}_{it}})^2 \tag{2}$$

where *k* represents a particular group of *K* groups, *I* is the number of participants, and T_i is the number of observations (per participant). The predicted value, $\hat{\boldsymbol{y}}_{it}$, for *M* variables at time *t* ($t > 1$) for participant *i* is given by the estimated VAR model for Group *k*. This procedure provides as many prediction errors as the number of groups for each participant; e.g., with two groups, two VAR models are fitted to each participant, resulting in two prediction errors per person. Third, group membership is updated according to prediction errors, i.e., by re-assigning participants to the group where they had the minimum prediction error. These steps are repeated to search for the most appropriate group membership, defined as the membership that minimizes the total prediction errors (i.e., L_k). Because prediction errors are evaluated on all the VAR models (not just on a univariate model), the ALS algorithm considers all variables in a model simultaneously to optimize the group partitioning. Note that the ALS algorithm does not explicitly address the hierarchical structure of data. Instead, it assumes a VAR model per group, whose parameters are homogenous within a group. Therefore, this approach is different from the multilevel VAR that has been used in the ESM/VAR literature (e.g., fitting a set of multilevel univariate models, allowing parameters vary across individuals even within a group).

To test the clustering accuracy of this algorithm, Bulteel et al. (2016) performed computer simulations varying parameters such as the number of observations, number of clusters, and size of cross-regression coefficients. The results indicated that the clustering accuracy increases as (a) group differences in the magnitude of regression coefficients increase, and (b) the number of observations (per participant) increases. Although the simulations of Bulteel et al. (2016) are already quite thorough, they assumed no individual differences in regression coefficients within a cluster. In their simulations, data were generated for a VAR model with 6 variables (i.e., 30 cross-regression coefficients to estimate). Each participant was given a homogenous set of regression coefficients within a cluster, but the population distributions to generate the coefficients could vary across clusters. For example, half of the coefficients in a model were sampled from a uniform distribution *U*[0.3, 0.5], whereas the other half were from *U*[0.0, 0.2]; and these sampled coefficients were the same across participants within a cluster. Although these parameter settings are consistent with the assumptions of the ALS algorithm (i.e., regression coefficients are identical across participants within a group), this assumption is somewhat too strict and it would be more natural to assume individual differences in each regression coefficient even within a group. More critically, this homogeneity assumption makes it difficult to scale the group differences that the ALS algorithm can detect, simply because the *SD*s of the regression coefficients are zero within a cluster. Furthermore, it is still unclear to what extent the ALS algorithm is sensitive to group differences that are present only in several coefficients of a VAR model. Given that a VAR model in affective dynamics contexts tends to be relatively large (e.g., a model with 10 variables has 100 potential auto- and cross-regression coefficients), it is very likely that not all (and even less than half of) the coefficients contribute to the clustering. One practical question is, therefore, what the minimum group distance (in terms of the number and size of group differences) is that can be identified by the ALS algorithm.

In the current study, we designed computer simulations to (a) examine the performances of ALS (compared to GMM) approaches for VAR-based clustering, and to (b) investigate how many "effective features" (EF; i.e., regression coefficients that vary across true clusters) should be included in a VAR model for the algorithms to identify the unknown clusters reliably. We predicted that it would be more difficult to predict group membership when the model has fewer effective features and/or more non-effective features. This is because non-effective features typically act as noise attenuating the influences of effective features in statistical classification (e.g., Manning, Raghavan, & Schütze, 2009). Thus, in Study 1, we explored the conditions where the ALS algorithm works reliably while systematically manipulating the number of EFs as well as the sizes of the group differences in the EFs. GMM was also evaluated under the same conditions. We started our computer simulation with an ideal situation, i.e., two groups with an equal sample size. Next, we used another mixing rate (i.e., the balance of the sample size between groups; Study 2a) and assumed three groups in data (Study 2b). Study 3 was a real data application; we demonstrated the ALS algorithm to analyze ESM data where negative affect was repeatedly assessed with an ESM design.

Study 1

Our computer simulations systematically manipulated (a) the number of variables in the model, (b) the number of effective features in the model, (c) the size of the group differences in regression coefficients. The ALS algorithms were applied to each dataset generated under these assumptions, which gave estimates of the clustering accuracy against the true (or simulated) group membership of participants. We also applied GMM on individualistically estimated VAR coefficients for the same simulation datasets, whose performance was compared to that of the ALS algorithm. Although the GMM approach appears to be theoretically more appropriate in a hierarchical dataset (i.e., allowing parameters vary across individuals even within a group), the individualistic VAR estimation might have more estimation errors than the ALS algorithm (Bulteel et al., 2016). Furthermore, GMM could be vulnerable to noise features that do not contribute to the clustering. Therefore, we predicted that the ALS algorithm would outperform the GMM approach in identifying the correct number of groups and in predicting group membership of participants.

Method

Data Generation. We manipulated the above mentioned three factors systematically, resulting in 2 × 3 × 3 conditions:

1. Number of variables (5 or 10) in a VAR model
2. Moderate, large, and very large effect sizes for group differences in regression coefficients: Cohen's $d = 0.5$, 1.0, and 2.0
3. Number of effective features (EFs; cross-regression coefficients that vary across groups): 1, 5, and 9 coefficients

Each simulation dataset consisted of 100 participants with 50 assessment occasions per participant. These are typical settings of ESM studies, e.g., with 10 beeps per day for seven days with a compliance rate of 80% (e.g, Ebner-Priemer & Trull, 2009; van Berkel, Ferreira, & Kostakos, 2017). To simplify the problem, two groups with an equal number of participants (50 vs. 50) were considered in the current simulations (cf. Study 2). The data generation processes were as follows: first, group-level **Φ** matrices were specified for the control and target group (CG and TG) in each condition. For example, in the 5-variable condition of EF = 1 with the very large effect size, group-level **Φ** matrices were:

$$\boldsymbol{\Phi}_{CG} = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.4 \end{bmatrix} \tag{3}$$

$$\boldsymbol{\Phi}_{TG} = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.4 \end{bmatrix} \tag{4}$$

The non-zero element (0.2) represents the EF. The EF(s) were randomly allocated in non-diagonal elements (i.e., cross-regression coefficients). Second, person-specific regression coefficients ($\boldsymbol{\Phi}_i$) were generated from these group-level **Φ** matrices. Autoregressive coefficients had no group differences, which were sampled from the same population distribution of *N*(0.40, 0.01) across all participants. Half of the participants (i.e., the control group) had cross-regression coefficients drawn from *N*(0.00, 0.01), whereas the other half (i.e., the target group) had coefficients sampled from *N*(0.05, 0.01), *N*(0.10, 0.01), and *N*(0.20, 0.01) for the conditions of moderate, large, and very large effect sizes, respectively. Therefore, the group differences (per cross-regression coefficient) are interpreted as Cohen's *d* = 0.5, 1.0, and 2.0. The number of these EFs in a VAR model was also manipulated systematically; for each of the three effect-size conditions, different numbers of cross-regression coefficients (EF = 1, 5, 9) were sampled from normal distributions with non-zero means, and the rest of the coefficients followed the zero-mean normal distribution that is the same as the control. For example, in the EF = 1 condition, 50 participants had a cross-regression coefficient that was sampled from *N*(0.20, 0.01) for the very large size group difference; the other cross-regression coefficients were from the same distribution as the control, i.e., *N*(0.00, 0.01). This procedure was repeated for each condition of effect sizes and numbers of variables in the model.

Using the sampled individual auto- and cross-regression coefficients, occasion-level data (i.e., responses at each ESM beep) were generated for each participant in the VAR(1) framework. The innovations (i.e., errors for outcome variables; ω) were sampled from *N*(0,1), and were assumed to be independent of each other. The intercept of each variable followed *N*(0.00, 0.01) with a very small covariance of 0.001. These intercept settings were used because it is common practice in ESM studies to center momentary variables by their person-means (so the intercepts should be distributed around zero and mostly independent across variables). For each of the 18 simulation conditions, 100 datasets (i.e., 10,000 participants) were generated to obtain distributions of the clustering performances of the clustering algorithms. The R code used for these simulations is available from OSF: https://osf.io/he4c8/

Stability assumption. A VAR process is stable if all eigenvalues of $\mathbf{\Phi}$ have modulus less than 1 (Lütkepohl, 2005). Because our simulations generated a random $\mathbf{\Phi}$ for each participant, some participants could have an instable process when they had extreme values for auto- and/or cross-regression coefficients. We examined how often this violation took place under our simulation settings, which was actually at a negligible rate: 0.2 – 0.5% of all simulation runs even in the conditions of the largest effect size. Therefore, we did not use any corrections or rescaling techniques on $\mathbf{\Phi}$ for further simulations and analyses. Another reason for this decision was that rescaling $\mathbf{\Phi}$ (Bulteel et al., 2016) makes it difficult to interpret the effect sizes.

Clustering by the ALS algorithm. The task of the ALS algorithm is to identify the number of groups that have different features in the VAR structure, and to predict the membership of participants in the identified groups. The ALS algorithm fits a VAR(1) model for each of *k*-number groups. The regression coefficients were therefore fixed within a group but are allowed to vary across groups. The optimization starts with assigning participants to the *k* groups. The initial membership is determined by hierarchical clustering (Ward method) on individualistic regression coefficients obtained by fitting a VAR(1) model for each participant (with the ordinary least square estimator; Bulteel et al., 2016).

To optimize the group partitioning, the ALS algorithm evaluates a total sum of predictions errors (residuals) from the *K* VAR(1) models, which are sequentially estimated by updating the participants' group membership. At each step of the updating, prediction errors are computed for each participant for each VAR(1) model. If a participant has a larger prediction error with the model fitted on the current group than that on the other group(s), this participant is re-assigned to the other group with the minimum prediction error. For example, if two groups are assumed with a 5-variable model, a fixed-effect VAR(1) model is estimated for each group (here group memberships is tentatively assigned to each participant). This step produces two VAR(1) models (two 5-by-5 $\mathbf{\Phi}$ matrices), which are then fitted to all participants in order to compute person-specific prediction errors. Each participant has two prediction errors from the two VAR models, which are evaluated to update the group membership of this participant. The participant is (re)assigned to the group where they had the minimum prediction error. This process is repeated for all participants, and the updated group membership is further used in the next optimization routine. The optimization routine is repeated until the group membership no longer changes. Because this procedure is easily trapped in a local minimum, it is recommended that random initial group membership be used as well as the rational membership given by hierarchical clustering (Bulteel et al., 2016).

The number of groups is determined by the CHull procedure (Ceulemans & Kiers, 2006; Wilderjans, Ceulemans, & Meers, 2013), which returns the number of groups that achieves the largest relative information gain when adding one extra group. The loss function (*L*$_k$), defined as a sum of prediction errors of *k* VAR models, is evaluated to search the maximum scree test ratio *st*:

$$st_k = \frac{L_{k-1} - L_k}{L_k - L_{k+1}} \tag{5}$$

The value of *k* that maximizes the *st* (i.e., adding another group does not improve the fit of the model) is regarded as the number of groups that best explains the data. In the current study, we evaluated *k* from 2 to 4, and tested whether the CHull procedure indicates *k* = 2 correctly for each run of simulations.

Clustering by GGM. We first fitted a VAR(1) model to the simulated data for each participant, whose estimates (regression coefficients) were then used for GMM clustering. The VAR parameters were estimated by Ordinary Least Squares (OLS). The clustering was performed by the R package*,* mclust (Scrucca, Fop, Murphy, & Raftery, 2016), which considers variant mixture models with different distribution structure types, volumes, shapes, and orientations of the cluster ellipsoids. Therefore, the clusters could have variable geometric characteristics; e.g., a cluster can be wider and more tilted than the other clusters. The model parameterization as well as the number of clusters was selected using the Bayesian Information Criterion (BIC). In general, the BIC prefers a more parsimonious model with less complexity, e.g., smaller number of clusters (e.g., Vandekerckhove, Matzke, & Wagenmakers, 2019).

Results and Discussion

Accuracy Predicting the Number of Groups. First, we evaluated the accuracy with which the ALS algorithm with CHull procedure suggests the number of groups when a VAR(1) model consists of five variables. Under the condition of moderately sized group differences, the algorithm suggested the correct number of groups at a probability of around 60% for all EF conditions (Table 1). Although the accuracy was not substantially improved for large group differences, it almost reached saturation between EF = 5 and 9 for very large group differences (5 variables). The results for a VAR model with 10 variables showed a similar pattern; the accuracy was the highest for EF = 5 and 9 with very large size effects (76 and 90%, respectively). Overall the prediction accuracy is lower in the conditions with 10 than 5 variables. In the 10-variable condition, the accuracy reached 80% only when the data had 9 EFs with very large group differences. We repeated the same set of simulations for extra 100 runs (per condition), which replicated the similar pattern of the results (Table 1, values in parentheses). There seems to be an interaction between the number and size of EFs, both of which have to be high in order to find a good performance of the ALS algorithm. This, in turn, means that it is difficult for the ALS algorithm to identify the correct number of groups when the feature space has only one large size EF out of 25 or 100 regression coefficients. Such a small number of EF(s) would be easily buried among the other random non-effective features, which do not contribute to clustering. Also EFs with a smaller size group difference hardly contribute to the clustering, as the EFs were simulated to be a local subset of dimensions in the feature space.

GMM failed to indicate the correct number of groups in almost all simulation conditions. The BIC typically preferred the most parsimonious model, suggesting only one group in the simulated sample. We explored the potential causes of this error in identifying the number of groups (see the supplementary materials for details). The first challenge of this approach is that the individualistic estimates of VAR parameters contain some estimation errors. We found that the magnitude of the correlation between the simulated and estimated regression coefficients was typically only moderate (mean *r* = .55). The second challenge is that GMM is less sensitive to EFs that are local in the feature matrix. We performed further dimension reduction by factor analysis on the estimated individual regression coefficients. When the factor analysis identified a smaller number of factors (i.e., 25 regression coefficients were summarized into 1 or 2 factor scores), GMM achieved better accuracy identifying the number of groups and predicting participants' group membership. However, the factor analysis often suggested 3 or more factors (note that the extra factors do not contain any useful information for clustering), which resulted in misspecification of the mixture model by the BIC. GMM was thus not evaluated in the following sections for parameter recovery and accuracy predicting group membership.

Parameter Recovery. Root mean square errors (RMSEs) were computed for the estimated regression coefficients (with a greater value indicating a larger deviance from the true coefficients) only for the simulation runs where the correct number of groups was suggested. RMSEs were specified at the group level. For each simulation run, we first computed the group means of regression coefficients that were generated for each individual; second, we subtracted these group means of the simulated regression coefficients from those of the ALS estimation; third, these difference scores were root mean squared across regression coefficients within a group, so that a mean RMSE was obtained for each group in each simulation run. The mean RMSE ranged from 0.021 to 0.037, and was typically distributed below 0.05 for the simulations with five variables (Figure 1, Panel A). Given that the regression coefficients were generated from normal distributions with a fixed *SD* of 0.10, the precision of parameter estimation can be regarded as good. Values of RMSEs were also small enough for the simulations with 10 variables, with the means ranging from 0.022 to 0.030 (Figure 1, Panel B)[3].

- Figure 1 inserted here -

Accuracy Predicting Group Membership. Classification accuracy (ACC), which is defined as (hits + correct rejections) / the total number of participants, was tested for each run of simulations. We computed the ACC (a) for all simulation runs and (b) only for runs where the correct number of groups was suggested. When the CHull procedure suggested an incorrect number of (i.e., three or more) groups, ACC was defined as the sum of participants who were assigned to the largest group in the true-control and true-target groups; e.g., if 30 (out of 50) true control participants were assigned to Group 1 and if 30 (out of 50) true target participants were assigned to Group 2, then the accuracy is 0.60; all participants assigned to Group 3 are regarded as incorrect predictions.

Figure 2 shows the distributions of ACC for each condition of the simulations with five variables in a model. When moderate size group differences were assumed, ACC distributed around the chance level, with means of 0.48 – 0.49 and SDs of 0.09 – 0.10. With large and very large group differences, ACC increased as a function of the number of EFs. To achieve 80% accuracy, at least 5 cross-regression coefficients (out of 25) had to have very large differences between the two groups. This tendency was observed also for ACC computed only on the runs where the correct number of groups was indicated; that is, even when the algorithm determined

[3] RMSE was inversely related to the accuracy predicting participants' group membership. This is because (a) we evaluated RMSE only when the correct number of groups was indicated; (b) even if the indicated number of groups was correct, the ALS algorithm could make a wrong prediction on participants' group membership, particularly in the conditions of less EFs and smaller effect sizes. In this case, VAR models were specified on incorrect grouping, which resulted in increased RMSE.

the correct number of groups, the predicted membership was largely different from the true membership when the group differences were only moderate to large.

- Figure 2 inserted here -

In the simulations with 10 variables, the performance of the ALS algorithm was slightly lower than in those with 5 variables (Figure 3). The prediction accuracy was typically distributed between 0.40 and 0.60 under the conditions of moderate to large group differences. To achieve 80% accuracy, at least nine EFs (our of 100 regression coefficients) with very large size group differences are needed. Again, this tendency was unchanged if ACC was computed on runs where the correct number of groups was indicated (Panel B).

- Figure 3 inserted here -

In summary, the ALS clustering showed good performance when there were sufficient group differences in the regression coefficients between two groups (i.e., 5 – 9 EFs with a very lager effect size). On the other hand, GMM clustering on the individual VAR estimates did not show an acceptable performance in any conditions. Therefore, GMM was not considered for further simulations and analyses in Studies 2 and 3.

Study 2

Study 1 examined the performance of the ALS algorithm under the assumption of two balanced groups. Study 2 explored the algorithm's performance further by assuming unbalanced sample sizes between the target and control groups (Study 2a) or three groups (one control and two targets; Study 2b). We set up computer simulations parallel to Study 1, manipulating the number of EFs, the effect sizes of the EFs, and the number of variables in the model, to determine whether the ALS algorithm could maintain performance comparable to that of Study 1, particularly under the conditions of 5 – 9 EFs with the very large effect size.

Method

Simulation data were generated for the same 2 × 3 × 3 conditions as in Study 1: the number of variables (5 or 10); effect sizes for group differences in EFs (Cohen's *d* = 0.5, 1.0, and 2.0); and number of EFs (1, 5, and 9 coefficients). The only differences were that we used a mixing rate of 30:70 for the control and target groups in Study 2a, and we assumed three groups in Study 2b. Study 2a used the same data generation process as Study 1. Individual $\mathbf{\Phi}$ matrices were simulated by sampling (a) individual auto-regression coefficients from a group-level distribution of $N(0.40, 0.01)$; (b) individual non-effective cross-regression coefficients from $N(0.00, 0.01)$; and (c) individual effective cross-regression coefficients from $N(0.05, 0.01)$, $N(0.10, 0.01)$, and $N(0.20, 0.01)$ for the moderate, large, and very large effect size conditions, respectively.

In Study 2b, we generated data for 120 participants instead of 100, with $N = 40$ in each group. We assumed a control group whose cross-regression coefficients were sampled from $N(0.00, 0.01)$, and two target groups with a variable number and size of effective cross-regression coefficients according to the simulation conditions. Group-level $\mathbf{\Phi}$ matrices for the two target groups were created by randomly allocating EFs in their non-diagonal elements, with no overlap in EFs between the two target groups. Example group-level $\mathbf{\Phi}$ matrices for the target groups are indicted below (under the condition of five variables, five EFs, and the very large effect size):

$$\boldsymbol{\Phi}_{TG=1} = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.4 & 0.0 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.0 & 0.2 & 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{bmatrix} \quad (5)$$

$$\boldsymbol{\Phi}_{TG=2} = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.2 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.4 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.0 \\ 0.2 & 0.2 & 0.0 & 0.0 & 0.4 \end{bmatrix} \quad (6)$$

For each simulated dataset, the ALS algorithm was applied with the same optimization and model selection (i.e., CHull) procedures as in Study 1.

Performance evaluation in Studies 2a and 2b was the same as in Study 1 unless explicitly mentioned. In both Study 2a and Study 2b, accuracy was defined as the sum of participants who were correctly assigned to the control and target group. Because Study 2a had unbalanced control and target group sample sizes, we computed specificity and sensitivity as additional performance measures. Specificity was defined as the ratio of participants who were correctly identified as "target" relative to participants who were simulated to be in the target group; sensitivity was defined as the ratio of participants who were identified as "control" relative to participants who were simulated to be in the control group.

In Study 2b, we found that the CHull procedure often suggested an incorrect number of groups (mostly two instead of three). In such cases, accuracy was the sum of participants who were assigned to the largest cell per identified group. For example, the ALS algorithm indicated two groups with the following characteristics:

- Group 1: 20, 30, and 10 participants from the true Control, Target-1, and Target-2 groups, respectively
- Group 2: 20, 10, 30 participants from the three groups, respectively.

In this example, accuracy was computed as (30 + 30) / 120 = 0.50. When the CHull procedure suggested more than three groups, accuracy was calculated from the largest three groups, and participants assigned to the smallest group were regarded as incorrect predictions.

Results and Discussion

Study 2a. The performance of the ALS algorithm was slightly lower when sample sizes were not balanced between the control and target groups. The algorithm indicated the correct number of groups in 60 – 70% of simulation runs and showed the best performance under the conditions of 9 EFs with the very large effect size (for both 5 and 10 variables). The accuracy of predicting participants' group membership showed a parallel pattern (Figure 4). The ALS prediction achieved a mean accuracy of > 80% for EFs = 5 and 9 with the very large effect size in the 5-variable condition, and for EF = 9 with the very large effect size in the 10-variable condition. Because of the unbalanced sample sizes between the two groups, we also examined sensitivity and specificity (Table S3); both were the highest in the EF = 9 condition with the very large effect size (sensitivity = 0.93, specificity = 0.97 for 5 variables; sensitivity = 0.88, specificity = 0.77 for 10 variables). Mean RMSEs (in simulation runs where the correct number of groups was indicated) ranged from 0.024 – 0.038 across conditions, comparable to the results observed in Study 1.

- Figure 4 inserted here -

Study 2b. The ALS algorithm was less accurate in indicating the number of groups when three groups were included in the sample (Table 1). The CHull procedure erroneously indicated two groups as the best solution for 28 – 69 (and 50 – 88) simulation runs in the five- (and 10-) variable conditions with the very large effect size. Interestingly, even in these error cases, the two target groups were well separated but the control group was typically split and absorbed into the two target groups. For example, the algorithm assigned the 40 participants from one target group to Group 1, the 40 participants from the other target group to Group 2, half of the 40 participants from the control group to Group 1, and the other half to Group 2. This is not surprising, because the control group is in between the two target groups in terms of the distance of **Φ** matrices. Therefore, regardless of its low performance in identifying the true number of groups, the ALS algorithm showed relatively good accuracy in predicting participants' group membership. Accuracy was distributed around 80% in the condition of EF = 9 with the very large effect size (Figure 5). We did not analyze RMSEs because of the low accuracy of identifying the correct number of groups.

- Figure 5 inserted here -

In summary, Study 2 showed that the ALS algorithm maintains acceptable performance in predicting participants' group membership even when groups are not balanced (Study 2a) or there are three groups (Study 2b), if the groups have sufficient differences in **Φ** matrices. However, performance was lower than that of Study 1 overall, highlighted by the increased number of errors in indicating the number of groups when three groups were assumed. Critically, a group that is in between two other groups (i.e., the control group in Study 2b) is likely to be overlooked due to the conservativeness of the CHull procedure.

Study 3

We applied the ALS algorithm to a real ESM dataset in which participants made online reports of their current mood in daily life. Momentary mood ratings on five items were used for the ALS clustering. We were particularly interested in how many features (regression coefficients) contributed to the clustering, and how large the features' group differences were.

Method

Data. University students (*N* = 99) received 10 beeps per day for six consecutive days (part of the data have been published elsewhere: Iijima, Takano, & Tanno, 2017). The beep-to-beep interval was approximately 90 min and was pseudo-randomized with a margin of ±15 min. In response to each beep, participants rated their current mood according to five items: tense, restless, uneasy, anxious, and nervous (from the tense-anxiety scale of the Profile of Mood State questionnaire; McNair, Lorr, & Droppleman, 1992). Each item was rated on a 5-point scale (0 = *not at all*, 4 = *very much*). We excluded the first responses of each day to keep the time interval more or less constant. Participants with fewer than 16 observations (*N* = 14) were not included in the clustering. The final dataset included 85 participants with 2,893 total observations; a median of 33 observations per participant (range: 16 – 57). All participants provided written informed consent and the study protocol was approved by the ethics committee of the University of Tokyo.

Estimation. We first applied the ALS algorithm to cluster individuals on the basis of the VAR(1) model. All variables were person-mean centered to eliminate between-person variance. Second, to evaluate the size of group differences in the auto- and cross-regression coefficients,[4] we estimated univariate multilevel models where each mood rating at time *t* was predicted by (a) the lagged mood ratings at time *t*-1, (b) the ALS-assigned group memberships, and (c) their interactions (e.g., Pe et al., 2015). A within-person-level model was formulated as follows:

$$Mood_{mti} = \beta_{0i} + \sum_{m=1}^{M} \beta_{mi}\, Mood_{m(t-1)i} + r_{it} \qquad (7)$$

[4] Note that the ALS algorithm does not provide variance estimates for regression coefficients (because it assumes no individual differences within a group).

where a rating on mood item m of participant i at time t ($Mood_{1ti}$) was predicted by M mood ratings at time t-1 with an intercept, β_{0i}, and residual, r_{it}. The auto- and cross-regression coefficients, β_{mi}, were allowed to vary across groups and individuals. Therefore, the between-person-level models were given as follows:

$$\beta_{mi} = \gamma_{m0} + \gamma_{m1} Group_i + u_{mi} \tag{8}$$

A group difference in a regression coefficient, γ_{m1}, was then standardized by the standard deviation of the corresponding random effect, u_{mi}, which is comparable to the effect sizes manipulated in Studies 1 and 2 (Feingold, 2009). The multilevel models were estimated using restricted maximum likelihood estimation, implemented in the lme4 R package (Bates, Maechler, Bolker, & Walker, 2015).

Results and Discussion

The CHull procedure suggested two groups as the best partitioning with *st* values of 2.38, 1.43, and 1.34 for the two, three, and four groups, respectively. The algorithm allocated 58 participants to Group 1 and the other 27 to Group 2. Overall, Group 2 was characterized by higher prospective effects of uneasiness than Group 1 on the other anxious feelings (Table 2). The matrices of the estimated regression coefficients met the stability assumption. When individuals in Group 2 felt uneasy at one moment, they tended to experience increases in tension, restlessness, anxiety, and nervous feelings at the next moment compared to those in Group 1; in other words, uneasiness appeared to be a precursor of more general and intense anxious feelings in Group 2. A potential issue that should be noted is that the variance components of several random effects were estimated to be (close to) zero. This suggests that the regression coefficients do not have random individual differences, and thus should be fixed within a group. In this case, the standardized effects of the group differences in the regression coefficients diverge to infinity because of the zero denominators.

The current real data analysis identified eight regression coefficients with large or very large group differences (greater than 1, Table 2). The group partitioning given by the ALS algorithm can be seen as reliable, as our simulations suggested good performance when 5 – 9 EFs had large or very large group differences (Study 1, Study 2a). Table 2 also illustrates the VAR parameter estimates for Group 1, indicating no substantial differences in the estimates between OLS regression embedded in the ALS routine vs. post-hoc multilevel modeling. The VAR parameters ranged from -0.10 to 0.37 (for Group 1), which are similar to coefficient magnitudes assumed in Studies 1 and 2. These results suggest that our simulation settings were realistic and appropriate in terms of the number and size of EFs and $\mathbf{\Phi}$ matrices.

General Discussion

We evaluated the clustering algorithms in identifying unknown groups of participants who share the same or similar features in several coefficients of a VAR(1) model. Simulation data were generated by varying (a) the number of variables in the model, (b) the size of the group differences in the cross-regression coefficients (or effect size), and (c) the number of regression coefficients that code group differences (or effective features; EFs). The results of Study 1 suggest that the accuracy of the ALS clustering is influenced by all three of these factors; more specifically, to achieve a sufficient clustering accuracy (> 80%), a VAR model has to have at least (a) five EFs with very large group differences ($d = 2.0$) in a 5-variable model; or (b) nine EFs with very large group differences in a 10-variable model. Even though the ALS algorithm assumes homogeneity in VAR parameters within a group, it showed overall good performance in data with a hierarchical structure (i.e., regression coefficients vary across individuals even within a group) when there are sufficient group differences in a subset of the coefficients.

Similar results were found in less ideal conditions where we assumed unbalanced sample sizes between groups (Study 2a) and three groups in the sample (Study 2b). Although the performance of the ALS algorithm was somewhat attenuated under these conditions (particularly for the model with 10 variables), the accuracy predicting individual group membership was acceptable when there were 5 – 9 EFs with the very large effect size. However, we found that the CHull procedure is somewhat too conservative to add an extra group as this procedure often missed the third group in the three-group condition (Study 2b). The missed group was typically in-between the other two groups in term of the **Φ**-matrix distance. Therefore, users of the ALS algorithm should be aware that it is still possible that there is another unidentified group even if the CHull procedure indicated the two groups as the best solution.

Together with the ALS algorithm, we evaluated performance of GMM clustering based on the individual VAR estimates. This approach was, however, hardly able to find the correct number of groups as the BIC often indicated that there was only one group in the sample. As discussed in Study 1 (and in the supplementary file), GMM has mainly two challenges to overcome for improving the clustering performance. First, the individualistic estimation of the VAR(1) parameters has only a moderate precision, which inevitably affects the subsequent clustering performances. Second and more critically, GMM is not sensitive to a smaller number of EFs relative to the random features that do not contribute to the clustering. If external criteria (or teacher signals) are available, feature selection by supervised learning would improve the accuracy of the classification (Brodersen et al., 2014). Another possibility is to use a feature selection algorithm for GMM (Scrucca & Raftery, 2014) to find a locally optimal subset of features with group information in a dataset. Although we tried this algorithm in our simulations as well, it was computationally too demanding to obtain the results in a timely manner for our simulations.

Feature selection is also relevant for the ALS algorithm, even though our simulation results suggest that the ALS is more sensitive to local EFs than GMM. Non-effective features add random noise to the loss function that is evaluated in the ALS algorithm, which contaminates the optimization process and increases error in clustering. In an analysis of empirical data, it is advisable to select variables from researchers' a priori knowledge about the clusters (e.g., to only include relevant variables that are expected to reflect hypothesized group differences) rather than to put all assessed variables into a model in a completely exploratory manner. One question that may come up at this point is whether the ALS algorithm can be used in an undefined variable space (i.e., when researches are not sure which variables or which coefficients are to be targeted for clustering). In an unsupervised-learning context, where the correct group labels are unknown, the goal of clustering is typically to find "interesting" groups of participants. Although the operational definition of "interesting" varies across contexts and goals of analyses, there are several statistical criteria (e.g., scatter separability and maximum likelihood criteria) that can be used to select meaningful (or effective) features for clustering (Dy & Brodley, 2004; Scrucca & Raftery, 2014). However, such an automatic feature selection does not necessarily lead to "better" clustering, as the selection criteria may not reflect researchers' interests. In that sense, it would be important to rely on some external source of knowledge that can serve as a criterion of the goodness of clustering. Typically, psychopathology research aims to identify a group of people who are vulnerable to a particular disorder. In this context, the levels of the symptomatology or the diagnosis of the disorder could be used to evaluate the clustering outcomes. Furthermore, if the "correct" group labels (or external criteria) are known, supervised learning would help researchers to select effective features, i.e., to search for the best set of variables to predict the "correct" group membership.

Related to this point, dimension reduction may also improve the performance of ALS clustering (Bulteel, Tuerlinckx, Brose, & Ceulemans, in press). The results of the current simulations show that clusters can be identified by a few coefficients in a VAR model. In turn, this suggests that if a VAR model includes a number of EFs that represent the same psychological construct (which are typically highly correlated), these features may have too strong an influence on (or even bias) the clustering. A questionnaire often includes items that tap into the same psychological construct but have different wordings. If the ALS algorithm is applied to such items, the clustering outcomes may only reflect the single dimension of that psychological construct and would mask the other hidden groups that are potentially psychologically interesting.

A critical limitation of the ALS algorithm is that it is not sensitive to moderately sized group differences in regression coefficients. Furthermore, with this effect size, the algorithm's performance was not improved even when the number of EFs was increased. In other words, the ALS algorithm should be used when one can expect (very) large group differences in the VAR estimates. Bulteel et al. (2016) applied this algorithm to time series data on depression-related symptoms, and found quite large and consistent group differences in the explained variance for criterion variables. Similarly, our real data analysis (Study 3) identified large and very large group differences in eight regression coefficients with the five-variable model. As a post-hoc check, we fitted the multilevel models with the ALS-identified groups as a moderator. These models provided the estimates of (a) the group differences in each regression coefficient and (b) their *SD*s as variance components of the random effects. Users of the ALS can, thus, refer to the standardized group differences in order to see if their ALS-identified groups have sufficiently large group differences in a good number of regression coefficients. Our tentative suggestion is to regard 5 – 9 regression coefficients with large and very large size group differences as a threshold to find a stable clustering result, although this varies depending on the number of variables in a VAR model. If researchers found only moderate group differences after their ALS clustering, the predicted group membership should be interpreted carefully. In that case, a sensitivity analysis (e.g., leaving out a participant or a feature regression coefficient, and performing the analysis on the remaining sample) would help researchers to know if their results are stable, and to identify which participants or which variables contribute to this instability.

Our simulations have several limitations that should be taken into account when interpreting the results. First, we fixed the number of participants and occasions at 100 participants with 50 responses per participant. We argue that it is not easy (though not impossible) to collect more than 50 complete responses per participant in an ESM study, but the performance of the ALS algorithm is generally more reliable the more responses that are available (Bulteel et al., 2016). Second, the current simulations are limited to a standard VAR(1) model, and thus the results could change for other models and analytic schemes. That said, the ALS algorithm would be applicable to many cases as it evaluates prediction errors, which can be defined regardless of the shapes of the models and estimators. One potential issue is that a VAR model typically assumes that the analyzed time series are stationary processes (i.e., means and regression coefficients are constant over time). If this assumption is violated, one could consider preprocessing the data, e.g., differencing, detrending, and deseasoning the time series. Another option is to use a model that allows coefficients to vary across time (e.g., Bringmann, Ferrer, Hamaker, Borsboom, & Tuerlinckx, 2018; Commandeur & Koopman, 2007), although such a model is more complex and needs more research to use with the ALS algorithm.

Third, our evaluation for the accuracy predicting participants' group membership may have been too conservative. We counted participants who were predicted to be in the third group as incorrect predictions in the two-group simulations (Studies 1 and 2a), but the third group could be seen as "correct" in some cases. For example, the third group can be similar to either the other two groups in term of the estimated $\mathbf{\Phi}$ matrices; that means, the algorithm identifies Group A vs. B1 and B2 (Groups B1 and B2 are interpreted as subgroups of Group B). Although this interpretation would increase the accuracy levels, we did not implement this idea in our evaluation because (a) the number of groups that is indicated by the CHull procedure is not trivial information in a real data analysis; and (b) because we found that the third group was typically in-between the other two groups (cf. Study 2b) and it was not always apparent which group the third group should be counted in.

In conclusion, the ALS algorithm reliably identifies subgroups of participants on the basis of a VAR(1) model when 10–20% of the regression coefficients in the model have very large group differences. Given that a set of non-effective features contaminates the ALS clustering, variable selection from a priori knowledge would be appropriate. In psychopathology research, levels of symptomatology and diagnosis of a target disorder would be good criteria for the variable selection as well as for validation of the clustering.

Declaration of interest statement

The authors declare that there is no conflict of interest.

Acknowledgement

We thank Yudai Iijima for sharing the ESM data.

References

Ahn, W., Krawitz, A., Kim, W., Busmeyer, J. R., & Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, *4*(2), 95–110. https://doi.org/10.1037/a0020684.A

Asparouhov, T., Hamaker, E. L., Muthén, B., Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic Structural Equation Models Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., … Snippe, E. (2019). What Do Centrality Measures Measure in Psychological Networks ? What Do Centrality Measures Measure in Psychological Networks ? *Journal of Abnormal Psychology*, *128*, 892–903. https://doi.org/10.1037/abn0000446

Bringmann, L. F., Ferrer, E., Hamaker, E. L., Borsboom, D., & Tuerlinckx, F. (2018). Modeling Nonstationary Emotion Dynamics in Dyads using a Time-Varying Modeling Nonstationary Emotion Dynamics in Dyads using a Time-Varying Vector-Autoregressive Model. *Multivariate Behavioral Research*, *53*(3), 293–314. https://doi.org/10.1080/00273171.2018.1439722

Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine*, *45*, 747–757. https://doi.org/10.1017/S0033291714001809

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. https://doi.org/10.1177/1073191116645909

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. https://doi.org/10.1371/journal.pone.0060188

Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., & Stephan, K. E. (2014). NeuroImage : Clinical Dissecting psychiatric spectrum disorders by generative embedding. *NeruoImage: Clinical*, *4*, 98–111. https://doi.org/10.1016/j.nicl.2013.11.002

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (n.d.). Improved insight into and prediction of network dynamics by combining VAR and dimension reduction. *Multivariate Behavioral Research*.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Clustering Vector Autoregressive Models : Capturing Qualitative Differences in Within-Person Dynamics. *Frontiers in Psychology*, *7*, 1540. https://doi.org/10.3389/fpsyg.2016.01540

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*, 133–150. https://doi.org/10.1348/000711005X64817

Commandeur, J. J. F., & Koopman, S. J. (2007). *An introduction to state space time series analysis*. Oxford, UK: Oxford University Press.

D'Urso, P., Di Lallo, D., & Maharaj, E. A. (2013). Autoregressive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks. *Soft Computing*, *17*(1), 83–131. https://doi.org/10.1007/s00500-012-0905-6

Dy, J. G., & Brodley, C. E. (2004). Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research*, *5*, 845–889.

Ebner-priemer, U. W., & Trull, T. J. (2009). Ecological Momentary Assessment of Mood Disorders and Mood Ecological Momentary Assessment of Mood Disorders and Mood Dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Epskamp, S., Deserno, M. K., & Bringmann, L. F. (2019). mlVAR: Multi-level vector autoregression (R pack- age version 0.4).

Epskamp, S., Waldorp, L. J., Mõttus, R., Borsboom, D., Epskamp, S., Waldorp, L. J., … Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Ernst, A. F., Timmerman, M. E., Jeronimus, B. F., & Albers, C. J. (2019). Insight Into Individual

Differences in Emotion Dynamics With Clustering. *Assessment*. https://doi.org/10.1177/1073191119873714

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods*, *14*(1), 43–53. https://doi.org/10.1037/a0014699

Frühwirth-Schnatter, S., & Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, *26*(1), 78–89. https://doi.org/10.1198/073500107000000106

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

Iijima, Y., Takano, K., & Tanno, Y. (2017). Attentional bias and its association with anxious mood dynamics. *Emotion*, *18*, 723–735. https://doi.org/10.1037/emo0000338.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of Affective Instability in Ecological Momentary Assessment: Indices Using Successive Difference and Group Comparison via Multilevel Modeling. *Psychological Methods*, *13*(4), 354–375. https://doi.org/10.1037/a0014173

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. *Sociological Methods and Research*, *29*(3), 374–393.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58.

https://doi.org/10.1016/j.jmp.2016.03.007

Klippel, A., Viechtbauer, W., Reininghaus, U., Wigman, J., van Borkulo, C., MERGE, … Wichers, M. (2018). The Cascade of Stress : A Network Approach to Explore Differential Dynamics in Populations Varying in Risk for Psychosis. *Schizophrenia Bulletin*, *44*(2), 328–337. https://doi.org/10.1093/schbul/sbx037

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion (Washington, D.C.)*, *12*(2), 283–289. https://doi.org/10.1037/a0025046

Leisch, F. (2019). FlexMix : A General Framework for Finite Mixture Models and Latent Class Regression in R. Retrieved from https://cran.r-project.org/web/packages/flexmix/vignettes/flexmix-intro.pdf

Liao, T. W. (2005). Clustering of time series data — a survey. *Pattern Recognition*, *38*, 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *Journal of Mathematical Psychology*, *55*(1), 68–83. https://doi.org/10.1016/j.jmp.2010.08.004

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

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1992). *Revised manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Services.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., … Gotlib, I. H. (2015). Emotion-Network Density in Major Depressive Disorder. *Clinical Psychological Science*, *3*(2), 297–300. https://doi.org/10.1177/2167702614540645

Pe, M. L., & Kuppens, P. (2012). The Dynamic Interplay Between Emotions in Daily Life : Augmentation , Blunting , and the Role of Appraisal Overlap. *Emotion*, *12*(6), 1320–1328. https://doi.org/10.1037/a0028262

Schuurman, N., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). 3 How to Compare Cross-Lagged Associations in a Multilevel Autoregressive Model. *Psychological Method*, *21*(2), 206–221. https://doi.org/10.1037/met0000062

Scrucca, L., Fop, M., Murphy, B. T., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J*, *8*(1), 289–317. Retrieved from https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf

Scrucca, L., & Raftery, A. E. (2014). clustvarsel : A Package Implementing Variable Selection for Model-based Clustering in R. Retrieved October 24, 2019, from https://arxiv.org/abs/1411.0606

van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, *50*(6), 93. https://doi.org/10.1145/3123988

Vandekerckhove, J., Matzke, D., & Wagenmakers, E. (2015). Model Comparison and the Principle of Parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 1–39). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

Wichers, M., Groot, P. C., & Group, E. S. M. (2016). Critical Slowing Down as a Personalized Early, 114–116. https://doi.org/10.1159/000441458

Wichers, M., Schreuder, M. J., Goekoop, R., & Groen, R. N. (2019). Can we predict the direction of sudden shifts in symptoms ? Transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological Medicine*, *49*(3), 380–387.

Wigman, J. T. W., Os, J. Van, Borsboom, D., Wardenaar, K. J., Epskamp, S., & Klippel, A. (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. https://doi.org/10.1017/S0033291715000331

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull : A generic convex-hull-based model selection method. *Behavior Research Methods*, *45*(1), 1–15. https://doi.org/10.3758/s13428-012-0238-5

Zheng, Y., Wiebe, R. P., Cleveland, H. H., Molenaar, P. C., & Harris, K. S. (2013). An Idiographic Examination of Day-to-Day Patterns of Substance Use Craving, Negative Affect and Tobacco Use among Young Adults in Recovery. *Multivariate Behavioral Research*, *48*(2), 241–266. https://doi.org/10.1080/00273171.2013.763012.An

Table 1

The Number of Simulation Runs Where the Number of Groups was Correctly Identified (Out of 100 Runs)

Study / Algorithm: ES conditions	With 5 variables			With 10 variables		
	EF =1	EF = 5	EF = 9	EF =1	EF = 5	EF = 9
Study 1: two groups, balanced mixing rate						
ALS: Moderate ES	62 (67)	50 (58)	62 (62)	60 (59)	64 (70)	79 (58)
ALS: Large ES	54 (64)	67 (59)	68 (70)	59 (65)	57 (68)	64 (61)
ALS: Very large ES	56 (46)	98 (96)	92 (98)	72 (59)	76 (74)	90 (89)
GMM: Moderate ES	0	0	0	0	0	0
GMM: Large ES	0	0	0	0	0	0
GMM: Very large ES	0	0	6	0	0	0
Study 2a: two groups, unbalanced mixing rate						
ALS: Moderate ES	50	54	66	63	65	68
ALS: Large ES	63	58	65	69	60	69
ALS: Very large ES	60	88	100	62	69	78
Study 2b: three groups, balanced mixing rate						
ALS: Moderate ES	28	19	24	27	19	15
ALS: Large ES	33	20	29	26	17	7
ALS: Very large ES	27	17	62	21	21	39

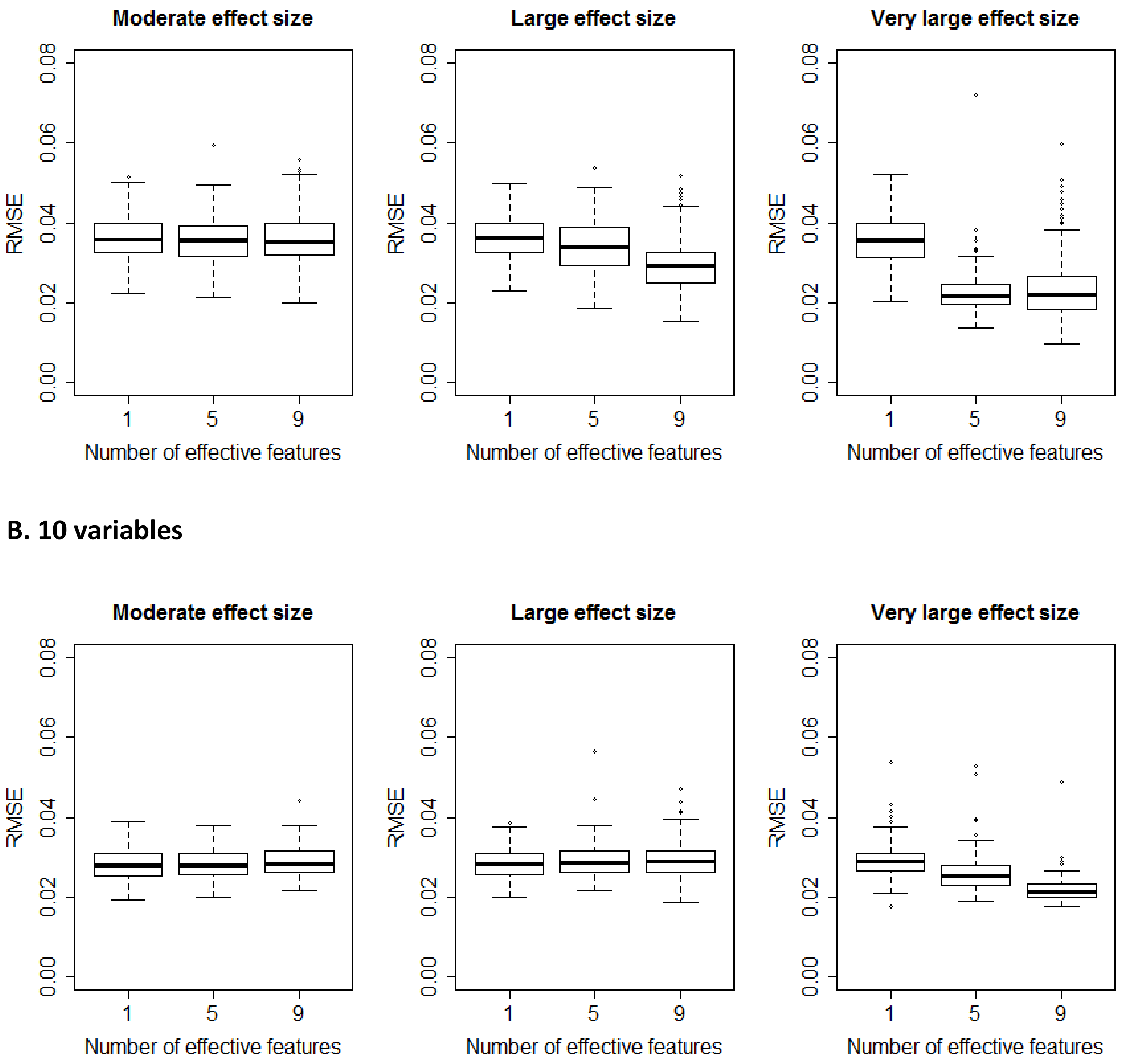
Note. In Study 1, we repeated all the simulations in order to test the stability of the results. The performances of the ALS algorithm in the second runs are indicated in the parentheses. ALS = Alternating Least Square algorithm; GMM = Gaussian Mixture Modeling; EF = Effective Features; ES = Effect size.

Table 2

Estimated Regression Coefficients and their Differences between the ALS-identified Groups

		ALS (G1)		Multilevel (G1)		Group differences (G1 – G2)		
DV	IV	Estimates	SE	Estimates (γ_{m0})	SE	Estimates (γ_{m1})	SD random effect (σ_{umi})	Standardized effect
Tense	Tense	0.19	0.05	0.17	0.05	-0.06	0.11	-0.51
Tense	Restless	-0.01	0.04	0.00	0.05	0.05	0.11	0.44
Tense	Uneasy	0.13	0.04	0.16	0.04	-0.21	0.00	Inf
Tense	Anxious	0.13	0.04	0.13	0.05	-0.12	0.15	-0.78
Tense	Nervous	0.06	0.04	0.06	0.04	-0.03	0.11	-0.26
Restless	Tense	0.04	0.05	0.07	0.05	-0.02	0.11	-0.18
Restless	Restless	0.09	0.04	0.06	0.05	0.02	0.11	0.20
Restless	Uneasy	0.14	0.05	0.13	0.05	-0.21	0.07	-3.10
Restless	Anxious	0.14	0.05	0.16	0.05	-0.10	0.14	-0.75
Restless	Nervous	0.10	0.04	0.09	0.04	-0.06	0.04	-1.54
Uneasy	Tense	-0.04	0.05	-0.01	0.05	0.08	0.09	0.83
Uneasy	Restless	0.03	0.05	0.02	0.05	-0.02	0.10	-0.17
Uneasy	Uneasy	0.38	0.05	0.37	0.04	-0.43	0.00	Inf
Uneasy	Anxious	0.09	0.05	0.10	0.05	-0.06	0.12	-0.54
Uneasy	Nervous	0.15	0.05	0.14	0.04	-0.02	0.06	-0.37
Anxious	Tense	0.15	0.05	0.12	0.04	-0.06	0.03	-1.73
Anxious	Restless	-0.13	0.04	-0.10	0.05	0.15	0.12	1.29
Anxious	Uneasy	0.21	0.04	0.21	0.04	-0.28	0.05	-5.91
Anxious	Anxious	0.21	0.04	0.16	0.05	-0.06	0.12	-0.53
Anxious	Nervous	0.04	0.04	0.02	0.04	0.01	0.12	0.11
Nervous	Tense	-0.04	0.05	-0.01	0.06	0.11	0.13	0.84
Nervous	Restless	0.02	0.05	0.01	0.05	-0.03	0.09	-0.37
Nervous	Uneasy	0.29	0.05	0.29	0.04	-0.34	0.02	-19.31
Nervous	Anxious	0.03	0.05	0.04	0.06	-0.09	0.17	-0.53
Nervous	Nervous	0.24	0.05	0.21	0.05	-0.09	0.11	-0.76

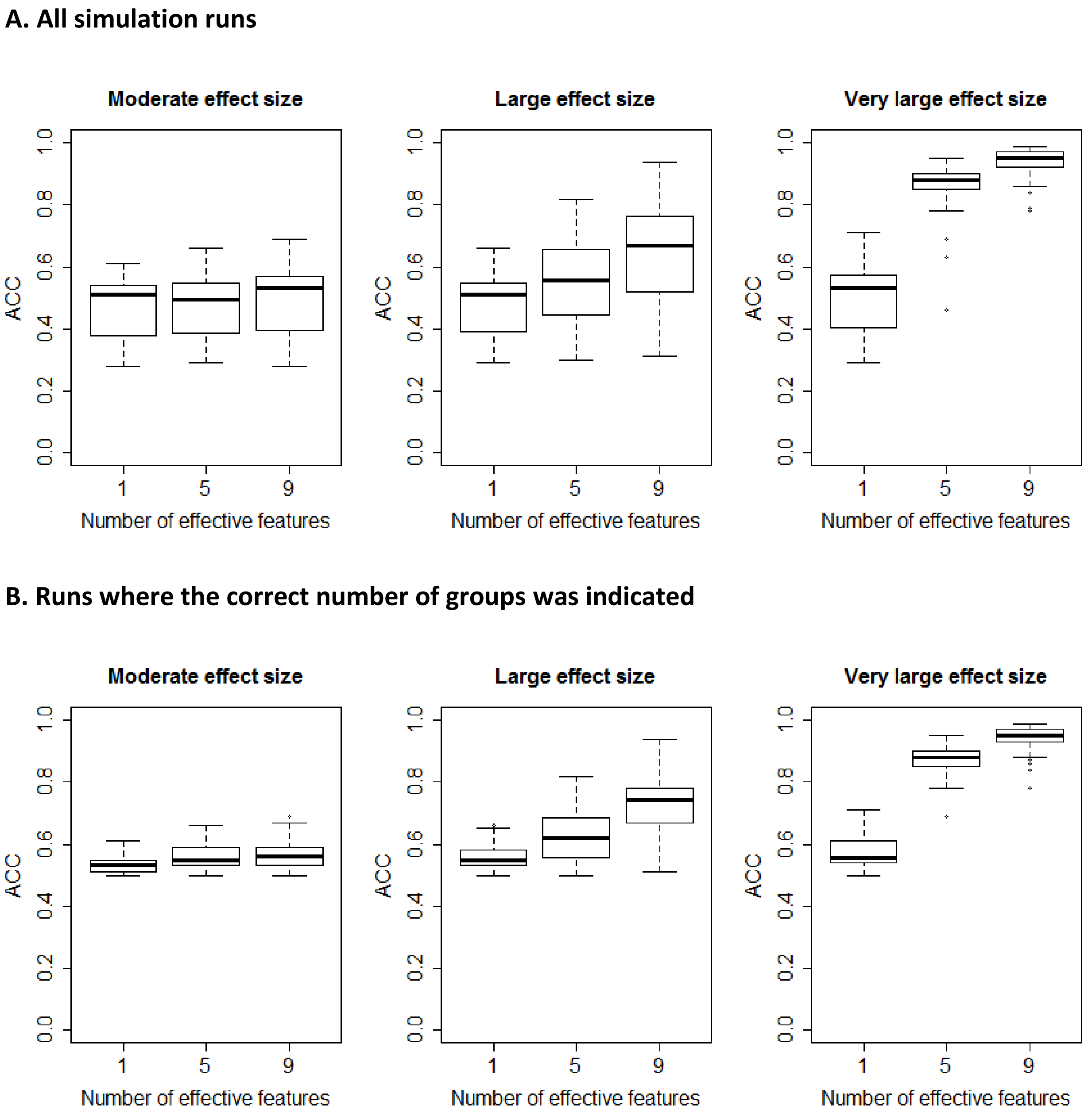
Note. ALS (G1) = regression coefficients (and SEs) for Group 1 estimated by the ALS algorithm, in which a VAR(1) was fitted on each group by OLS; Multilevel (G1) = regression coefficients (and SEs) for Group 1 estimated by multilevel modeling;



A. 5 variables

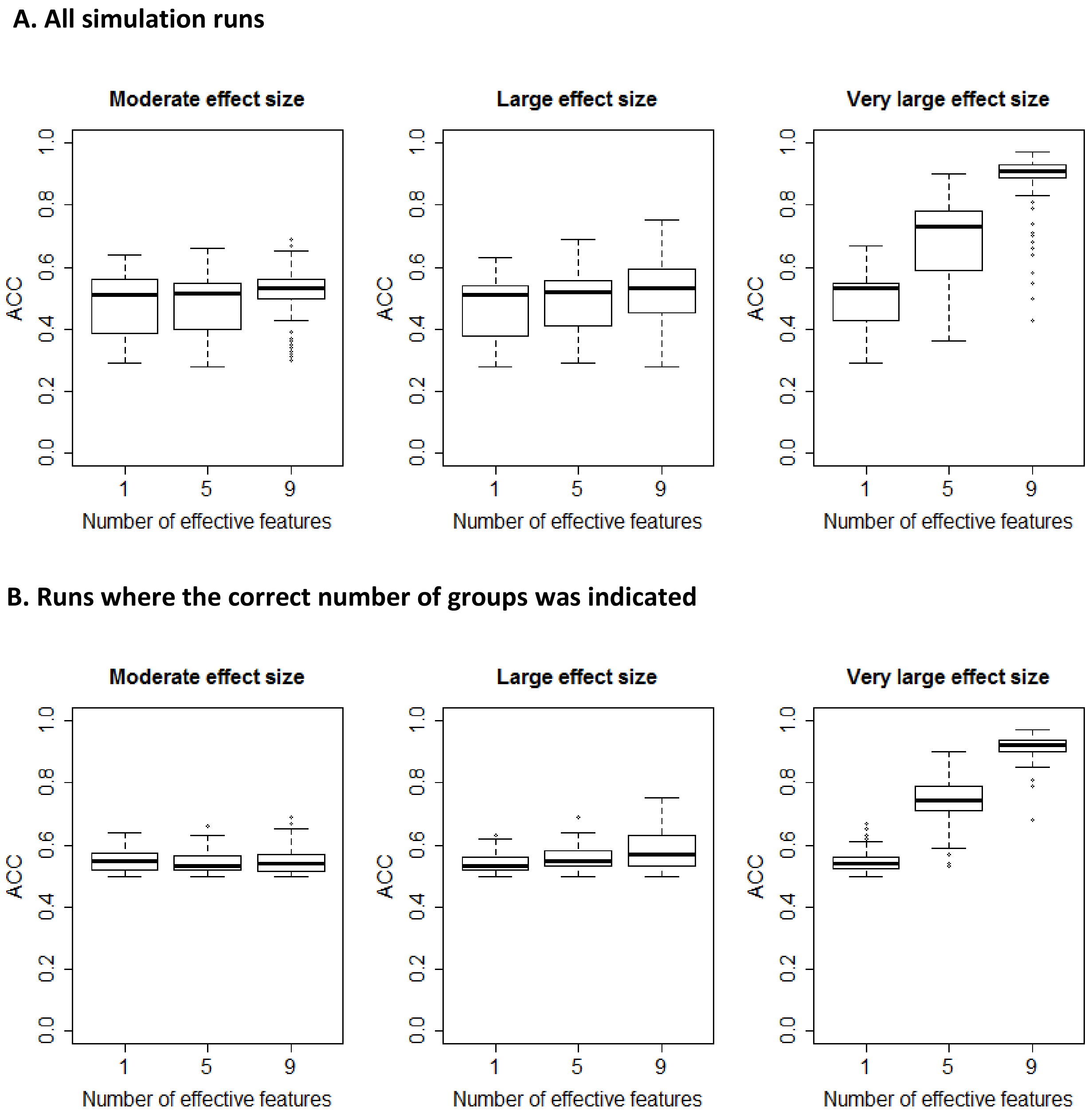
B. 10 variables

Figure 1. Root mean square errors (RMSE) of estimated regressive coefficients as a function of number of effective features for variable effect sizes (Study 1). Each point indicates a mean RMSE across regression coefficients per group, per simulation run. Panel A: Simulations with 5 variables in a model. Panel B: Simulations with 10 variables in a model.


A. All simulation runs

B. Runs where the correct number of groups was indicated

Figure 2. Classification accuracy (ACC) of the alternating least square algorithm for a vector autoregressive model with 5 variables (Study 1). Data were simulated with variable effect sizes and number of effective features in a model. Panel A: accuracy for all simulation runs. Panel B: accuracy for runs in which the correct number of groups was suggested.



A. All simulation runs

B. Runs where the correct number of groups was indicated

Figure 3. Classification accuracy (ACC) of the alternating least square algorithm for a vector autoregressive model with 10 variables (Study 1). Data were simulated with variable effect sizes and number of effective features in a model. Panel A: accuracy for all simulation runs. Panel B: accuracy for runs in which the correct number of groups was suggested.

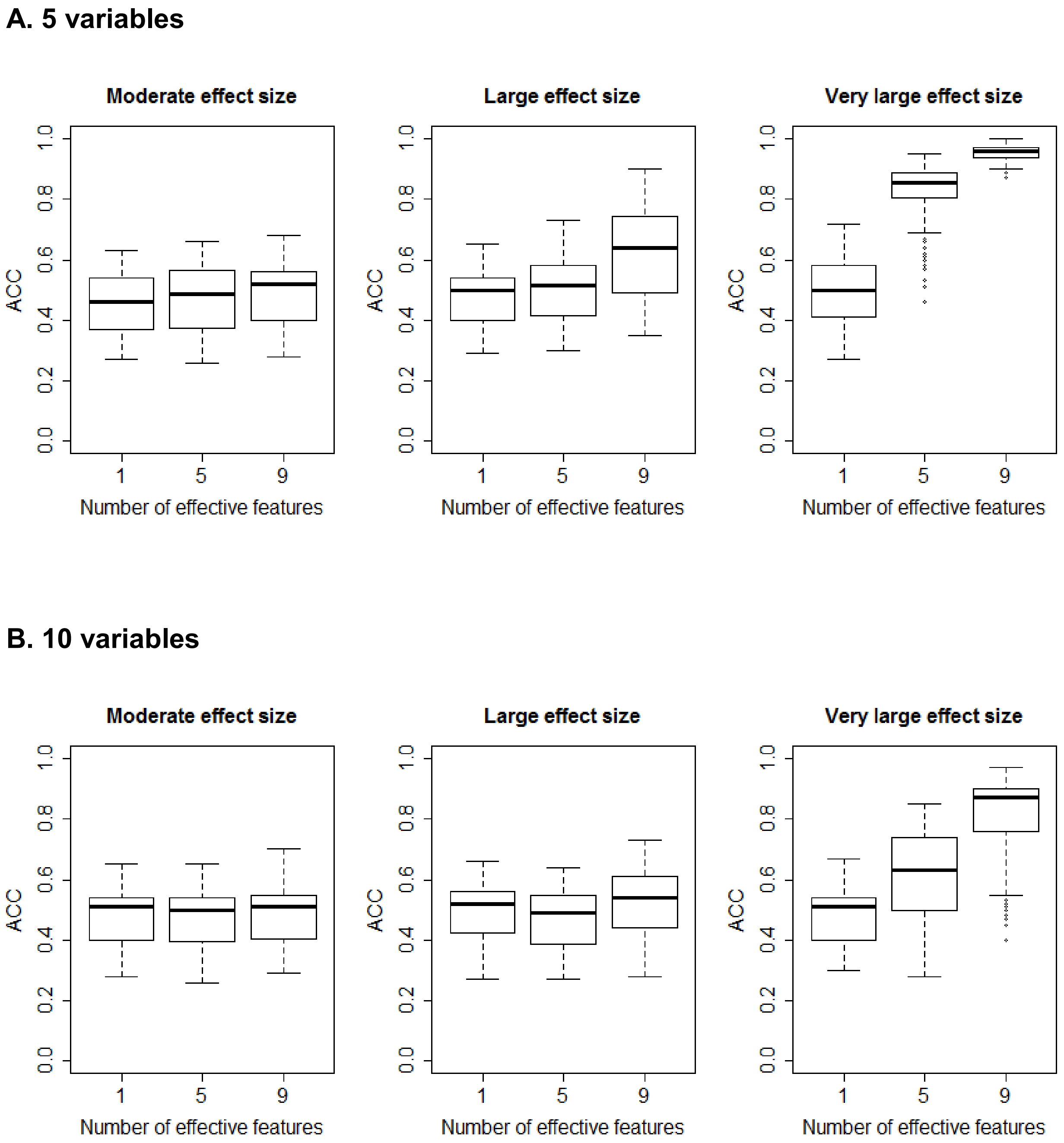


Figure 4. Classification accuracy (ACC) of the alternating least square algorithm when unbalanced sample sizes were assumed (Study 2a). All simulation runs are included.

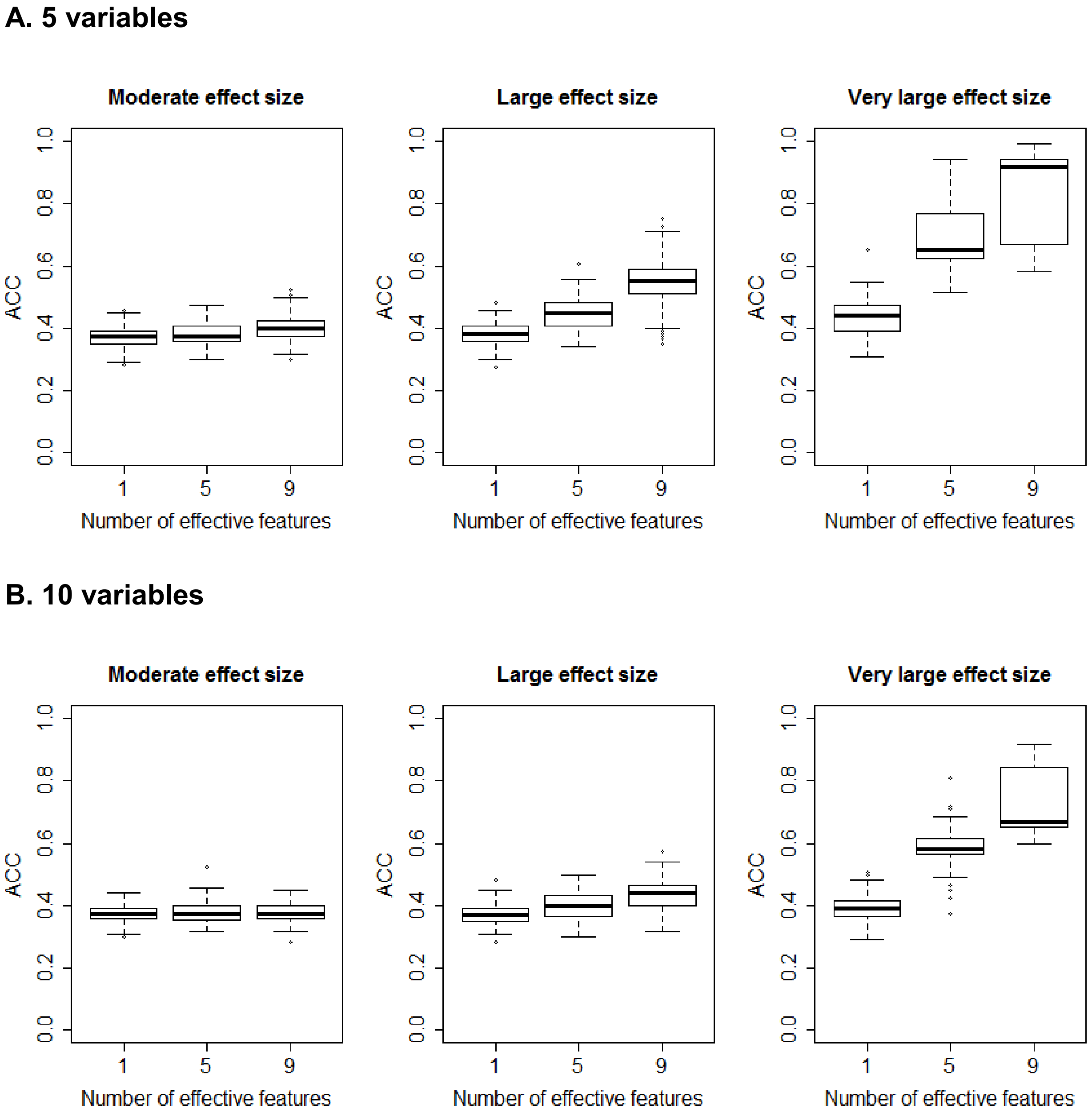


Figure 5. Classification accuracy (ACC) of the alternating least square algorithm when three groups were assumed (Study 2b). All simulation runs are included.