Mixture Simultaneous Factor Analysis for Capturing Differences in Latent Variables Between Higher-level Units of Multilevel Data

Kim De Roover
KU Leuven, Tilburg University

Jeroen K. Vermunt
Tilburg University

Marieke E. Timmerman
University of Groningen

Eva Ceulemans
KU Leuven

Author Notes:

Kim De Roover is a post-doctoral fellow of the Fund for Scientific Research Flanders (Belgium). The research leading to the results reported in this paper was sponsored in part by Belgian Federal Science Policy within the framework of the Interuniversity Attraction Poles program (IAP/P7/06), by the Research Council of KU Leuven (GOA/15/003), and by the Netherlands Organization for Scientific Research (NWO) [Veni grant 451-16-004]. Correspondence concerning this paper should be addressed to Kim De Roover, Quantitative Psychology and Individual Differences Research Group, Tiensestraat 102, B-3000 Leuven, Belgium. E-mail: Kim.DeRoover@kuleuven.be.

Abstract

Given multivariate data, many research questions pertain to the covariance structure: whether and how the variables (for example, personality measures) covary. Exploratory factor analysis (EFA) is often used to look for latent variables that may explain the covariances among variables; for example, the Big Five personality structure. In case of multilevel data, one may wonder whether or not the same covariance (factor) structure holds for each so-called 'data block' (containing data of one higher-level unit). For instance, is the Big Five personality structure found in each country or do cross-cultural differences exist? The well-known multigroup EFA framework falls short in answering such questions, especially for numerous groups/blocks. We introduce mixture simultaneous factor analysis (MSFA), performing a mixture model clustering of data blocks, based on their factor structure. A simulation study shows excellent results with respect to parameter recovery and an empirical example is included to illustrate the value of MSFA.

Keywords: factor analysis, mixture model clustering, multilevel data, latent variables

1. Introduction

Given multivariate data, researchers often wonder whether the variables covary to some extent and in what way. For instance, in personality psychology, there has been a debate about the structure of personality measures (i.e., the 'Big Five' versus 'Big Three' debate; De Raad et al., 2010). Similarly, emotion psychologists have discussed intensely whether and how emotions as well as norms for experiencing emotions can be meaningfully organized in a low-dimensional space (e.g., Ekman, 1999; Fontaine, Scherer, Roesch, & Ellsworth, 2007; Russel & Barrett, 1999; Stearns, 1994). Factor analysis (Lawley & Maxwell, 1962) is an important tool in these debates as it explains the covariance structure of the variables by means of a few latent variables, called factors. When the researchers have a priori assumptions on the number and nature of the underlying latent variables, confirmatory factor analysis (CFA) is often used, whereas exploratory factor analysis (EFA) is applied when one has no such assumptions.

Research questions about the covariance structure get further ramifications when the data have a multilevel structure; for instance, when personality measures are available for inhabitants from different countries. We will refer to data organized according to the higher level units (e.g., the countries) as 'data blocks'. For multilevel data, one can wonder whether or not the same structure holds for each data block. For example, is the Big Five personality structure found in each country or not (De Raad et al., 2010)? Similarly, many cross-cultural psychologists argue that the structure of emotions and emotion norms differ between cultures (Eid & Diener, 2001; Fontaine, Poortinga, Setiadi, & Markam, 2002; MacKinnon & Keating, 1989; Rodriguez & Church, 2003).

When looking for differences and similarities in covariance structures, using EFA is very advantageous because it leaves more room for finding differences than CFA does (see Section 2.3). For instance, in the emotion norm example (Eid & Diener, 2001), one may very well expect two latent variables to show up in each country corresponding to approved and

disapproved emotions, while being clueless about which emotions will be (dis)approved and how this differs across countries. In search for such differences and similarities, one may perform a multigroup or multilevel[1] EFA (Dolan, Oort, Stoel, & Wicherts, 2009; Hessen, Dolan, & Wicherts, 2006; Muthén, 1991), or an EFA per data block. These methods fall short in answering the research question at hand, however. Multigroup/multilevel EFA can be used to test whether or not between-group differences in factors are present, but neither of them indicate how they are different and for which data blocks. When multigroup/multilevel EFA indicates the presence of between-block differences, one can compare the block-specific EFA models to pinpoint differences and similarities. But when many groups are involved, the numerous pairwise comparisons are neither practical nor insightful; i.e., it is hard to draw overall conclusions based on a multitude of pairwise similarities and dissimilarities. For instance, in Section 4, we present data on emotion norms for 48 countries. Since multigroup EFA indicates that the factor structure is not equal across groups, comparing the group-specific structures would be the next step. It would be a daunting task, however, with no less than 1128 pairwise comparisons. More importantly, subgroups of data blocks may exist that share essentially the same structure and finding these subgroups is substantively interesting. Multilevel mixture factor analysis (MLMFA; Varriale & Vermunt, 2012) performs a mixture clustering of the data blocks based on some parameters of their underlying factor model, but it does not allow the factors themselves to differ across the data blocks.

Within the deterministic modeling framework however, a method exists that clusters data blocks based on their underlying covariance structure and performs a simultaneous component analysis (SCA, which is a multigroup extension of standard principal component

1 Note that multilevel EFA (Muthén, 1991) models the pooled within-block covariance structure and the covariance structure of the block-specific means by lower- and higher-level factors, respectively. A connection between equality of the lower- versus higher-order factor structure and invariance of within-block factors across data blocks has been shown (Jak, Oort, & Dolan, 2013), however.

analysis; Timmerman & Kiers, 2003) per cluster. The so-called ‘clusterwise SCA’ (De Roover, Ceulemans, & Timmerman, 2012; De Roover, Ceulemans, Timmerman, Nezlek, & Onghena, 2013; De Roover, Ceulemans, Timmerman, & Onghena, 2013; De Roover et al., 2012) has proven its merit in answering questions pertaining to differences and similarities in covariance structures (Brose, De Roover, Ceulemans, & Kuppens, 2015; Krysinska, et al., 2014). However, the method also has an important drawback, which follows from its deterministic nature, in that no inferential tools are provided for examining parameter uncertainty (e.g., standard errors, confidence intervals), conducting hypothesis tests (e.g., to determine which factor loading differences between clusters are significant), and performing model selection. Furthermore, even though similarities between component and factor analyses have been well-documented (Ogasawara, 2000; Velicer & Jackson, 1990; Velicer, Peacock, & Jackson, 1982), the theoretical status of components and factors is not the same (Borsboom, Mellenbergh, & van Heerden, 2003; Gorsuch, 1990). Therefore, to examine covariance structure differences in terms of differences in underlying latent variables (i.e., unobservable variables that have a causal relationship to the observed variables), such as the above-mentioned personality traits and affect dimensions, an EFA-based method is to be preferred.

Therefore, we introduce mixture simultaneous factor analysis (MSFA), which encompasses a mixture model clustering of the data blocks, based on their underlying factor structure. MSFA can be estimated by means of Latent GOLD (LG; Vermunt & Magidson, 2013) or Mplus (Muthén & Muthén, 2005). Even though the stochastic framework provides many inferential tools, various adaptations of the software will be necessary to reach the full inferential potential of the MSFA method (i.e., for the tools to be applicable for MSFA, as will be explained later on). Therefore, this paper focuses mainly on the model specification and an extensive evaluation of the goodness-of-recovery, i.e., how well MSFA recovers the clustering as well as the cluster-specific factor models.

The remainder of this paper is organized as follows: In Section 2, the multilevel multivariate data structure and its preprocessing is discussed, as well as the model specifications of MSFA, followed by its model estimation and its relations to existing mixture and/or multilevel factor analysis methods. In Section 3, the performance of MSFA is evaluated in an extensive simulation study. Section 4 illustrates the method with an application. Finally, Section 5 concludes the paper with points of discussion and directions for future research.

2. Mixture Simultaneous Factor Analysis

2.1. Data Structure and Preprocessing

We assume multilevel data, which implies that observations or lower-level units are nested within higher-level units (e.g., patients within hospitals, pupils within schools, inhabitants within countries). Both the lower- and the higher-level units are assumed to be a random sample of the population of lower- and higher-level units, respectively. We will index the higher-level units by $i = 1, \ldots, I$ and the lower-level units by $n_i = 1, \ldots, N_i$. The data of each higher-level unit i is gathered in a $N_i \times J$ data matrix or 'data block' $\mathbf{X}_i$, where J denotes the number of variables. Since MSFA focuses on modeling the covariance structure of the data blocks (within-block structure; Muthén, 1991), irrespective of differences and similarities in their mean level (between-block structure), all data blocks are columnwise centered before the analysis.

2.2. Model Specification

MSFA applies common factor analysis at the observation level and a mixture model at the level of the data blocks. Specifically, we assume (1) that the observations are sampled from a mixture of normal distributions that differ with respect to their covariance matrices, but all

have a zero mean vector (which corresponds to all data blocks being columnwise centered beforehand, see Section 2.1[2]), and (2) that all observations of a data block are sampled from the same normal distribution.

More formally, the MSFA model can be written as follows:

$$f\left(\mathbf{X}_i;\boldsymbol{\theta}\right)=\sum_{k=1}^{K}\pi_k f_k\left(\mathbf{X}_i;\boldsymbol{\theta}_k\right)=\sum_{k=1}^{K}\pi_k\prod_{n_i=1}^{N_i}MVN(\mathbf{x}_{n_i};\Sigma_k)\quad with\quad \boldsymbol{\Sigma}_k=\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k{}'+\mathbf{D}_k \qquad (1)$$

where f is the total population density function, and $\boldsymbol{\theta}$ refers to the total set of parameters. Similarly, f_k refers to the kth cluster-specific density function and $\boldsymbol{\theta}_k$ refers to the corresponding set of parameters. The latter densities are specified as K normal distributions, the covariance matrices of which are modeled by cluster-specific factor models. Thus, $\boldsymbol{\theta}_k$ refers to the cluster-specific factor loadings in the $J \times Q$ matrix $\boldsymbol{\Lambda}_k$ (implying the number of factors Q to be the same across clusters[3]) and the unique variances on the diagonal of $\mathbf{D}_k$. The mixing proportions (i.e., the prior probabilities of a data block belonging to each of the clusters) are indicated by π_k, with $\sum_{k=1}^{K}\pi_k=1$. Equation 1 implies the following additional assumptions: Firstly, the cluster-specific covariance matrices are perfectly modeled by the corresponding low-rank cluster-specific factor models (i.e., no residual covariances, implying that $\mathbf{D}_k$ is a diagonal matrix). Secondly, within each block, the observations are locally independent, warranting the use of the multiplication operator in Equation 1. Thirdly, we impose the factor scores and the residuals to be normally distributed for each data block, with the covariance matrix of the factor scores being an identity matrix and that of the residuals being equal to $\mathbf{D}_k$. In this paper, the

[2] An alternative would be to include block-specific (rather than cluster-specific) means in the model (see Section 5). This does not affect the obtained solution.

[3] Allowing for a different number of factors across the clusters complicates the comparison of cluster-specific models and implies a severe model selection problem (e.g., De Roover, Ceulemans, Timmerman, Nezlek, & Onghena, 2013) that needs to be scrutinized in future research.

factor (co)variance matrix is restricted to equal identity for each data block, in order to capture all differences in observed-variable covariances by means of the cluster-specific factor loadings – which implies creating the exact stochastic counterpart of the clusterwise SCA variant described by De Roover and colleagues (2012). This has the interpretational advantage of establishing all structural differences without having to inspect the (possibly many) block-specific factor (co)variances. Of course, more flexible model specifications in terms of the factor (co)variances are possible. Note that the cluster-specific factors have rotational freedom, which we take into account by using a rotational criterion, such as VARIMAX (Kaiser, 1958) or generalized Procrustes rotation (Kiers, 1997) that enhances the interpretability of the factor loading structures. Because factor rotation is not yet included in LG, we take the loadings estimated by LG 5.1 and rotate them in Matlab R2015b.

By means of Bayes' theorem, the posterior classification probabilities of the data blocks can be calculated, giving information regarding the blocks' cluster memberships and the uncertainty about this clustering. Specifically, these probabilities pertain to the posterior distribution (i.e., conditional on the observed data) of the latent cluster memberships z_{ik}:

$$\gamma\left(z_{ik}\right) = f\left(z_{ik}=1 \mid \mathbf{X}_i;\boldsymbol{\theta}\right) = \frac{f\left(\mathbf{X}_i, z_{ik}=1\right)}{f\left(\mathbf{X}_i\right)} = \frac{\pi_k f_k\left(\mathbf{X}_i;\boldsymbol{\theta}_k\right)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}\left(\mathbf{X}_i;\boldsymbol{\theta}_{k'}\right)} \tag{2}$$

2.3. Relations to existing methods

Since MSFA is an exploratory method, we omit related confirmatory methods like mixture factor analysis (Lubke & Muthén, 2005; Muthén, 1989; Yung, 1997), factor mixture analysis (Blafield, 1980; Yung, 1997), multilevel factor mixture modeling (Kim, Joo, Lee, Wang, & Stark, 2016), and a number of multigroup CFA extensions (Asparouhov & Muthén, 2014; Jöreskog, 1971; Muthén & Asparouhov, 2013; Sörbom, 1974). As mentioned in the

Introduction, methods based on CFA leave less room to find differences. Indeed, CFA imposes an assumed structure of zero-loadings upon the factors; thus, CFA-based methods can only account for differences in the size of the freely estimated (i.e., non-zero) factor loadings. Specifically, we compare MSFA to (1) a non-multilevel mixture EFA model, called 'mixtures of factor analyzers' (MoFA; McLachlan & Peel, 2000), and (2) a multilevel mixture EFA model: MLMFA (Varriale & Vermunt, 2012).

MoFA performs a mixture clustering of individual observations based on their underlying EFA model. The observation-level clusters differ with respect to their intercepts, factor loadings and unique variances, whereas the factors have means of zero and an identity covariance matrix per cluster. In contrast, MSFA deals with block-centered multilevel data and clusters data blocks (instead of individual observations) based on their factor loadings and unique variances (omitting the intercepts).

MLMFA models between-block differences in intercepts, factor means and unique variances by a mixture clustering of the data blocks, but MLMFA requires equal factor loadings across blocks. Hence, the MLMFA model specification differs in the following respects from MSFA. Firstly, unlike in MSFA, the cluster-specific means of the K multivariate normal distributions are not restricted to zero and capture between-block differences in mean levels on either the observed variables (intercepts) or the latent variables (factor means). Secondly, unlike MSFA, MLMFA models differences in covariance structures by means of differences in unique variances and factor (co)variances but not by differences in factor loadings (i.e., in contrast to Equation 1, loadings are common across clusters). Thus the range of covariance differences that MLMFA can capture is rather limited when compared to MSFA. Moreover, since both mean levels and covariance structures are taken into account, the MLMFA clustering will often be dominated by the means because they have a larger influence on the fit, whereas with MSFA mean differences are discarded.

2.4. Model Estimation

The unknown parameters $\boldsymbol{\theta}$ of the MSFA model are estimated by means of maximum likelihood (ML) estimation. This involves maximizing the logarithm of the likelihood function:

$$\begin{aligned}\log L(\boldsymbol{\theta} \mid \mathbf{X}) &= \log\left(\prod_{i=1}^{I}\sum_{k=1}^{K}\pi_k\prod_{n_i=1}^{N_i}\frac{1}{(2\pi)^{J/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}'\right)\right)\\ &= \sum_{i=1}^{I}\log\left(\sum_{k=1}^{K}\pi_k\prod_{n_i=1}^{N_i}\frac{1}{(2\pi)^{J/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}'\right)\right).\end{aligned} \tag{3}$$

where $\mathbf{X}$ is the $N \times J$ data matrix – with $N = \sum_{i=1}^{I} N_i$ – that is obtained by vertically concatenating the I data blocks $\mathbf{X}_i$. Note that the likelihood function is computed as a product of the likelihood contributions of the I data blocks, assuming that they are a random sample and thus mutually independent. To find the parameter estimates $\hat{\boldsymbol{\theta}}$ that maximize Equation 3, maximum likelihood estimation is performed by LG, which uses a combination of an EM algorithm and a Newton-Raphson algorithm (see Appendix 1). Because the standard random starting values procedure turned out to be rather prone to local maxima (especially when the number of clusters or factors increases), we experimented with alternative starting procedures. Appendix 1 describes the procedure we used, which involves starting with a PCA solution to which randomness is added.

3. Simulation Study

3.1. Problem

To evaluate the model estimation performance in terms of the sensitivity to local maxima and goodness of recovery, data sets were generated (by LG 5.1) from an MSFA model with a known number of clusters K and factors Q. We manipulated six factors that all affect

cluster separation: (1) the between-cluster similarity of factor loadings, (2) the number of data blocks, (3) the number of observations per data block, (4) the number of underlying clusters and (5) factors, and (6) between-variable differences in unique variances. Factor 1 pertains to the similarity of the clusters and we anticipate the performance to be lower when clusters have more similar factor loadings. Factors 2 and 3 pertain to sample size. We expect the MSFA algorithm to perform better with increasing sample sizes (i.e., more data blocks and/or observations per data block) (de Winter*, Dodou* & Wieringa, 2009; Steinley & Brusco, 2011). With respect to Factors 4 and 5, i.e., the complexity of the underlying model, we hypothesize that the performance will decrease with increasing complexity (de Winter*, Dodou* & Wieringa, 2009; Steinley & Brusco, 2011). Factor 6, between-variable differences in unique variances, was manipulated to study whether and to what extent the performance of MSFA is affected by these differences. Theoretically, MSFA should be able to deal with these differences perfectly (Section 2.2), in contrast to the existing clusterwise SCA which makes no distinction between common and unique variances (De Roover et al. 2012).

3.2. Design and Procedure

The six factors were systematically varied in a complete factorial design:

1. the *between-cluster similarity of factor loadings* at 2 levels: medium, high similarity;
2. the *number of data blocks I* at 3 levels: 20, 100, 500;
3. the *number of observations per data block* N_i at 5 levels: for the sake of simplicity, N_i is chosen to be the same for all data blocks; specifically, equal to 5, 10, 20, 40, 80;
4. the *number of clusters K* at 2 levels: 2, 4;
5. the *number of factors Q* at 2 levels: 2, 4;
6. *between-variable differences in unique variances*: differences among the diagonal elements in $\mathbf{D}_k$ ($k = 1, \ldots, K$) are either absent or present (explained below);

With respect to the cluster-specific factor loadings, a binary simple structure matrix was used as a common base for each $\mathbf{\Lambda}_k$. In this base matrix, the variables are equally divided over the factors, i.e., each factor gets six loadings equal to one in the case of two factors, and three loadings equal to one in case of four factors (see Table 1). To obtain medium between-cluster similarity (factor 1), each cluster-specific loading matrix $\mathbf{\Lambda}_k$ was derived from this base matrix by shifting the high loading to another factor for two variables, whereas these variables differ among the clusters (see Table 1). For the high similarity level, each $\mathbf{\Lambda}_k$ was constructed from the base matrix by adding, for each of two variables, a crossloading of $\sqrt{(.4)}$ and lowering the primary loading accordingly (see Table 1). Note that the factor loadings are constructed such that each observed variable has the same common variance per cluster – i.e., $(1 - e_k)$, where e_k is the mean of the unique variances within a cluster. To quantify the similarity of the obtained cluster-specific factor loading matrices, they were orthogonally Procrustes rotated to each other (i.e., for each pair of $\mathbf{\Lambda}_k$ matrices, one was chosen to be the target matrix and the other was rotated towards the target matrix) and a congruence coefficient φ (Tucker, 1951) was computed[4] for each pair of corresponding factors in all pairs of $\mathbf{\Lambda}_k$ matrices, where a congruence of one indicates that the two factors are proportionally identical. Subsequently, a grand mean of the obtained φ-values was calculated, over the factors and cluster pairs. On average, φ amounted to .73 for the medium similarity condition and .93 for the high similarity condition.

[Insert Table 1 about here]

[4] The congruence coefficient (Tucker, 1951) between two column vectors **x** and **y** is defined as their normalized inner product: $\varphi_{\mathbf{xy}} = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}}\sqrt{\mathbf{y}'\mathbf{y}}}$.

Regarding factor 6, the first level of this factor was realized by simply setting each diagonal element of $\mathbf{D}_k$ equal to e_k. For the second level, differences in unique variance were introduced by ascribing a unique variance of ($e_k - e_k/2$) to a randomly chosen half of the variables and a unique variance of ($e_k + e_k/2$) to the other half of the variables.

The simulated data were generated as follows: The number of variables J was fixed at 12 and an overall unique variance ratio e of .40 was pursued for all simulated data sets, where $e = \frac{1}{JK}\sum_{k=1}^{K} trace(\mathbf{D}_k) = \frac{1}{K}\sum_{k=1}^{K} e_k$. Between-cluster differences in e_k were introduced for all data sets, because they are usually present in empirical data sets. Specifically, in case of two clusters, the e_k values are .20 and .60, whereas in case of four clusters, the intermediate values of .30 and .50 are added for the additional clusters. To keep the overall variance equal across the clusters, the $\mathbf{\Lambda}_k$ matrices were rowwise rescaled by $\sqrt{1-e_k}$. Finally, to make the simulation more challenging, the cluster sizes were made unequal. Specifically, the data blocks are divided over the clusters such that one cluster is three times smaller than the other cluster(s). Thus, in case of two clusters, 25% of the data blocks were in one cluster and 75% in the other one. In case of four clusters, the small cluster contained 10% of the data blocks whereas the other clusters consisted of 30% each. The cluster memberships were generated by randomly assigning the correct number of data blocks to each cluster, according to these cluster sizes.

For each cell of the factorial design, 20 raw data matrices $\mathbf{X}^r$ were generated, using the LG simulation procedure, as described in Appendix 3. The $\mathbf{X}_i^r$ matrices resulting from the procedure were centered per variable, and their vertical concatenation yields the total data matrix **X**. In total, 2 (between-cluster similarity of factor loadings) × 3 (number of data blocks) × 5 (number of observations per data block) × 2 (number of clusters) × 2 (number of factors) × 2 (between-variable differences in unique variances) × 20 (replicates) = 4,800 simulated data

matrices were generated. Each data matrix **X** was analyzed by means of a LG syntax specifying an MSFA model with the correct number of clusters *K* and factors *Q* (e.g., Appendix 2) and applying 25 different sets of initial values (generated as described in Appendix 1). No convergence problems were encountered in this simulation study.

3.3. Results

First, the sensitivity to local maxima is evaluated. Secondly, the goodness of recovery is discussed for the different aspects of the MSFA model: the clustering, the cluster-specific factor loadings, and the cluster-specific unique variances. Thirdly, as an instance of the inferential tools provided by LG, the standard errors of the parameter estimates are assessed. Finally, some overall conclusions are drawn.

3.3.1. Sensitivity to local maxima

To evaluate the occurrence of local maximum solutions, we should compare the log *L* value of the best solution obtained by the multistart procedure with the global ML solution for each simulated data set. The global maximum is unknown, however, because the simulated data do not perfectly comply with the MSFA assumptions and contain error. Alternatively, we make use of a 'proxy' of the global ML solution; i.e., the solution that is obtained when the algorithm applies the true parameter values as starting values. The final solution from the multistart procedure is then considered to be a local maximum when its log *L* value is smaller than the one from the proxy. By this definition, only 227 (4.7%) local maxima were detected over all 4,800 simulated data sets. Not surprisingly, most of these occur in the more difficult conditions; e.g., 179 of the 227 local maxima are found in the conditions with a high between-cluster similarity of the factor loadings and 153 are found for the most complex models, i.e., when *K* as well as *Q* equal four.

3.3.2. Goodness of cluster recovery

To examine the goodness of recovery of the cluster memberships of the data blocks, we will (1) compare the modal clustering (i.e., assigning each data block to the cluster for which the posterior probability is the highest) to the true clustering, and (2) investigate the degree of certainty of these classifications. To compare the modal clustering to the true one, the *Adjusted Rand Index* (*ARI*; Hubert & Arabie, 1985) is computed. The *ARI* equals 1 if the two partitions are identical, and equals 0 when the overlap between the two partitions is at chance level. The mean *ARI* over all data sets amounts to .93 (*SD* = 0.18), which indicates a good recovery. The *ARI* was affected most by the amount of available information. Specifically, the mean *ARI* for the conditions with only 20 data blocks and five observations per block was only .51, whereas the mean over the other conditions was .96.

To examine the 'classification certainty' (CC), we computed the following statistics:

$$CC_{mean} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} \hat{z}_{ik}\gamma\left(z_{ik}\right)}{I} \quad \text{and} \quad CC_{min} = \min_{i} \sum_{k=1}^{K} \hat{z}_{ik}\gamma\left(z_{ik}\right) \tag{4}$$

where $\gamma\left(z_{ik}\right)$ and $\hat{z}_{ik}$ indicate the posterior probabilities (Equation 2) and the modal cluster memberships (i.e., estimates of the latent cluster membership z_{ik}), respectively. On average, CC_{mean} and CC_{min} amount to .9983 (*SD* = 0.007) and .94 (*SD* = 0.14), respectively, indicating a very high degree of certainty for the simulated data sets. Because CC_{mean} hardly varies over the simulated data sets, we focused on CC_{min} and inspected to what extent it is related to cluster recovery. To this end, a scatter plot of CC_{min} versus the *ARI* is given in Figure 1. From this figure, it is apparent that lack of classification certainty often does not coincide with classification error or the other way around.

3.3.3. Goodness of loading recovery

To evaluate the recovery of the cluster-specific loading matrices, we obtained a goodness-of-cluster-loading-recovery statistic (*GOCL*) by computing congruence coefficients φ (Tucker, 1951) between the loadings of the true and estimated factors and averaging across factors and clusters as follows:

$$GOCL = \frac{\sum_{k=1}^{K}\sum_{q=1}^{Q}\varphi\left(\boldsymbol{\lambda}_{kq}, \hat{\boldsymbol{\lambda}}_{kq}\right)}{KQ} \tag{5}$$

with $\boldsymbol{\lambda}_{kq}$ and $\hat{\boldsymbol{\lambda}}_{kq}$ indicating the true and estimated loading vector of the *q*-th factor for cluster *k*, respectively. The rotational freedom of the factors per cluster was dealt with by an orthogonal procrustes rotation of the estimated towards the true loading matrices. To account for the permutational freedom of the cluster labels, the permutation was chosen that maximizes the *GOCL* value. The *GOCL* statistic takes values between 0 (no recovery at all) and 1 (perfect recovery). For the simulation, the average *GOCL* is .98 (*SD* = 0.04), which corresponds to an excellent recovery. As for the clustering, the loading recovery depends strongly on the amount of information; i.e., the mean *GOCL* is .87 for the conditions with only 20 data blocks and five observations per block and .99 for the remaining conditions.

3.3.4. Goodness of unique variance recovery

To quantify how well the cluster-specific unique variances are recovered, we calculated the mean-absolute-difference (*MAD*) between the true and estimated unique variances:

$$MAD_{uniq} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}\left|d_{kj} - \hat{d}_{kj}\right|}{KJ} \tag{6}$$

On average, the MAD_{uniq} was equal to .06 (*SD* = 0.06). Like the cluster and loading recovery, the unique variance recovery depends most on the amount of information; i.e, the MAD_{uniq} has a mean value of .22 for the conditions with 20 data blocks or five observations each and .05 for

the other conditions. Also, the MAD_{uniq} value is affected by the occurrence of Heywood cases (Van Driel, 1978), a common issue in factor analysis pertaining to 'improper' factor solutions with at least one unique variance estimated as being negative or equal to zero. When this occurs during the estimation process, LG restricts it to be equal to a very small number (Vermunt & Magidson, 2013). Therefore, for the simulation, we consider a solution to be a Heywood case when at least one unique variance in one cluster is smaller than .0001. This was the case for 633 (13.2%) out of the 4,800 data sets, most of which occurred in the conditions with 20 blocks or five observations per block and thus with small within-cluster sample sizes (i.e., 601 out of the 633), or in case of four factors per cluster (i.e., 522 out of the 633). Specifically, the mean MAD_{uniq} is equal to .18 for the Heywood cases and .04 for the other cases. In the literature, a Heywood case has been considered a diagnostic of problems such as (empirically) underdetermined factors or insufficient sample size (McDonald & Krane, 1979; Rindskopf, 1984; Van Driel, 1978; Velicer & Fava, 1998).

3.4. Conclusion

The low sensitivity to local maxima indicated that the applied multistart procedure is sufficient. The goodness-of-recovery for the clustering, and cluster-specific factor loadings and unique variances was very good, even in case of very subtle between-cluster differences in loading pattern, and was mostly affected by the within-cluster sample size.

4. Application

To illustrate the empirical value of MSFA, we applied it to cross-cultural data on norms for experienced emotions from the International College Survey (ICS) 2001 (Diener et al., 2001; Kuppens, Ceulemans, Timmerman, Diener, & Kim-Prieto, 2006). The ICS study included 10,018 participants out of 48 different nations. Each of them rated, among other things,

how much each of 13 emotions (listed in Table 3) is appropriate, valued and approved in their society, using a 9-point likert scale (1 = “people do not approve it at all”, 9 = “people approve it very much”). Participants with missing data were excluded, so that 8894 participants are retained. Differences between the countries in the mean norm ratings were removed by centering the ratings per country (see Section 2.1).

MSFA is applied to this data set to explore differences and similarities in the covariance structure of emotion norms across the countries. To this end, the number of clusters and factors to use needs to be specified. Within the stochastic framework of MSFA, different information criteria are readily available. Even though the BIC (Schwarz, 1978) is often used for factor analysis and/or clustering methods (Bulteel, Wilderjans, Tuerlinckx, & Ceulemans, 2013; Dziak, Coffman, Lanza, & Li, 2012; Fonseca & Cardoso, 2007), its performance for MSFA model selection still needs to be evaluated (see Section 5). Therefore, model selection is based on interpretability and parsimony for this empirical example.

With respect to the number of factors, we a priori expect a factor corresponding to the positive (i.e., approved) emotions and a factor corresponding to the negative (i.e., disapproved) emotions. To explore this hypothesis and to confirm the presence of factor loading differences, we performed multigroup EFA by means of the R packages Lavaan 0.5-15 and SemTools 0.4-0 (Rosseel, 2012). A multigroup EFA with group-specific loadings for one factor indicated a bad fit (CFI = .74, RMSEA = .14), whereas the fit for two (group-specific and orthogonal) factors was excellent (CFI = .99, RMSEA = .03) (Hu & Bentler, 1999); thus, supporting the hypothesis of two factors. When restricting the loadings of these two factors to be invariant across countries, the fit dropped severely (CFI = .78, RMSEA = .12). The latter confirms that the countries differ in their factor loadings and, thanks to MSFA, the 1128 pairwise comparisons across the 48 country-specific EFA models are no longer needed to explore these differences.

The comparison of MSFA models with different numbers of clusters and two factors per clusters indicated that, in general, the same two extreme factor structures were always found, with the additional clusters only leaving more room for setting apart data blocks with an 'intermediate' factor structure. Thus, we select the MSFA model with two clusters and two factors per cluster. The clustering of the selected model is given in Table 2. Most countries are assigned to the clusters with a posterior probability of 1, whereas a negligible amount of classification uncertainty is found for Slovakia and South Africa. In order to validate and interpret the obtained clusters, we looked into some demographic measures that were available on the countries. An interesting difference between the clusters pertained to the Human Development Index (HDI) 1998, which was available from the Human Development Report 2000 (United Nations Development Programme, 2000) for 47 out of the 48 countries in the ICS study (i.e., only lacking for Kuwait). The HDI takes on values between 0 and 1 and measures the average achievements in a country in terms of life expectancy, knowledge and a decent standard of living. Figure 2a depicts boxplots of the HDI per cluster and shows that Cluster 1 contains less developed countries. Another aspect distinguishing the clusters was the level of conservatism (Schwartz, 1994), which was available for half of the countries only. Conservatism corresponds to a country's emphasis on maintaining the status quo, propriety, and restraining actions or desires that may disrupt group solidarity or traditional order. Specifically, as Figure 2b shows, the countries in Cluster 1 are more conservative than the ones in Cluster 2.

To shed light on how the covariance structure of emotion norms differs between the conservative and less developed countries on the one hand and the progressive and developed countries on the other hand, we first look at the varimax rotated cluster-specific factor loading matrices in Table 3. As expected, the two factors correspond to positive/approved and negative/disapproved emotions, respectively, and they do so in both clusters, indicating that the

within-country covariance structures have much in common. In addition to slight differences in the size of primary and/or secondary loadings, the most important and interesting cross-cultural difference is found with respect to ‘pride’. Specifically, in Cluster 1, the primary loading of ‘pride’ is on the ‘negative’ factor, whereas, in Cluster 2, its primary loading is on the ‘positive’ factor. Thus, by applying MSFA, we conveniently learned that in the conservative and less developed countries of Cluster 1, pride is a disapproved emotion, while in the progressive, developed countries of Cluster 2, pride is more positively valued by society. Possibly, in Cluster 1 pride is considered to be an expression of arrogance and superiority, whereas in Cluster 2 it is regarded a sign of self-confidence, which is a valued trait in progressive and developed countries. To complete the picture of the covariance differences, the cluster-specific unique variances are given in Table 4. From this table, it is apparent that all items have a higher unique variance in Cluster 2, implying that they have more idiosyncratic variability in the progressive, developed countries.

In addition to the visual comparison of the cluster-specific loadings (and unique variances), hypothesis testing is useful to discern which differences are significant or not. By default, Latent GOLD provides the user with results of Wald tests for factor loading differences across clusters (Vermunt & Magidson, 2013). We need to eliminate the rotational freedom of the cluster-specific factors for these results to make sense, however (see Section 5). This can be done by a sensible set of loading restrictions such as echelon rotation (Dolan, Oort, Stoel, & Wicherts, 2009; McDonald, 1999) and choosing these restrictions is easier in case of less clusters and less factors per cluster. In Table 3 (above), we see that ‘jealousy’ has a loading of (almost) zero in both clusters. Restricting this loading to be exactly zero in both clusters imposes echelon rotation and solves the rotational freedom. The resulting clusters-specific loadings are given in the lower portion of Table 3 and they hardly differ (i.e., the difference is never larger than .03) from the Varimax rotated ones. As indicated by a ** or * behind the loadings (Table

3, below), 8 factor loadings are significantly different between the clusters at the 1% level, whereas 10 are significantly different at the 5% level (Bonferroni correction for multiple testing was applied)[5].

5. Discussion

In this paper, we presented mixture simultaneous factor analysis, a novel exploratory method for clustering groups (i.e., higher-level units or 'data blocks', in general) with respect to the underlying factor loading structure as well as their unique variances. When researchers have statistical, empirical or theoretical reasons to expect possible differences, MSFA provides a solution to evaluate which differences exist and for which blocks. The solution is parsimonious because of the clustering of the data blocks, implying that only a few cluster-specific factor loading matrices need to be compared to pinpoint the differences in factor structure. Moreover, the clustering is often an interesting result in itself.

In this paper, the MSFA model was specified as the exact stochastic counterpart of clusterwise SCA-ECP (De Roover et al., 2012), i.e., with block-specific factor (co)variance matrices equal to identity, such that all differences in observed-variable covariances are captured between the clusters, by their cluster-specific factor loading matrices. Of course, in some cases, more flexible specifications are preferable; for instance, when one wants data blocks with the same factors but different factor (co)variances to be assigned to the same cluster. Another alternative model specification may include block-specific intercepts, instead of using

[5] Note that Wald-test results are also available for differences in unique variances. For the emotion norm data set, all between-cluster differences in unique variances were significant at the 1% level.

data block centering, thus preserving information on block-specific mean levels and capturing them in the model.

In contrast to clusterwise SCA, MSFA provides information on classification uncertainty, when present. Also, common variance is distinguished from unique variance (including measurement error). Thus, in specific cases wherein the unique variances differ strongly between variables and/or between clusters, MSFA will capture the underlying latent structures and the corresponding clustering more accurately. When this is not the case, clusterwise SCA may give similar results.

Of course, when pursuing inferential conclusions, the stochastic framework is to be preferred. For instance, it may be interesting to look at the standard errors of the parameter estimates. More importantly, with respect to the factor loading differences, one may argue that visual comparison of the cluster-specific loadings is too subjective. Conveniently, hypothesis testing for factor loading differences is available within the stochastic framework of MSFA and in LG (see Section 4). As stated in the Introduction, these inferential tools are not yet readily applicable for MSFA, which is due to the rotational freedom of the cluster-specific factors. For now, for the standard errors and Wald test results to make sense, rotational freedom can be eliminated by imposing loading restrictions, as was illustrated in Section 4. To avoid this choice of restrictions and its possible influence on the clustering, standard error estimation should be combined with the specification of rotational criteria for the cluster-specific factor structures. It is important to note that the factor rotation of choice affects which differences are found between the clusters, be it visually or by means of hypothesis testing. Therefore, future research will include evaluating the influence and suitability of different rotational criteria. Rotational criteria pursuing both between-cluster agreement and simple structure of the loadings seem appropriate (Kiers, 1997; Lorenzo-Seva, Kiers, & ten Berge, 2002) and the criteria can be

converted into loading constraints to be imposed directly during maximum likelihood estimation (Archer & Jennrich, 1973; Jennrich 1973).

The rotational freedom per cluster is a consequence of our choice for an exploratory approach (i.e., using EFA instead of CFA per cluster). Developing an MSFA variant with CFA within the clusters might be interesting for very specific cases like imposing the Big Five structure of personality for one cluster and the Big Three for the other cluster (De Raad et al., 2010) to see which countries end up in which cluster. Note that a priori assumptions on the underlying factor structure(s) can also be useful for the current, exploratory MSFA, i.e., as a target structure when rotating the cluster-specific factor structures and when selecting the number of factors.

Finally, the obtained factor loading differences and clusters depend on the number of clusters as well as the number of factors within the clusters. Therefore, solving the so-called 'model selection problem' is imperative. To this end, the performance of the BIC for MSFA model selection will be thoroughly evaluated and adaptations will be explored if needed. The fact that the BIC performance needs to be scrutinized is illustrated by the fact that, for the application, the BIC selected seven clusters, which appears to be an overselection when comparing cluster-specific factors and considering the (lack of) interpretability and stability of the clustering. Adaptations that will be considered include the hierarchical BIC (Zhao, Jin, & Shi, 2015; Zhao, Yu, & Shi, 2013) and stepwise procedures like the one described by Lukočienė, Varriale and Vermunt (2010). Their performances will be investigated and compared, also for the more intricate case wherein the number of factors may vary across clusters.

References

Archer, C. O., & Jennrich, R. I. (1973). Standard errors for orthogonally rotated factor loadings. *Psychometrika, 38,* 581-592.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 495–508.

Battiti, R. (1992). First-and second-order methods for learning: between steepest descent and Newton's method. *Neural computation, 4,* 141–166.

Blafield, E. (1980). Clustering of observations from finite mixtures with structural information. *Jyvaskyla, Finland: Jyvaskyla University*.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Brose, A., De Roover, K., Ceulemans, E., & Kuppens, P. (2015). Older adults' affective experiences across 100 days are less variable and less complex than younger adults'. Psychology and Aging, 30, 194-208. doi:10.1037/a0038690

Bulteel, K., Wilderjans, T. F., Tuerlinckx, F., & Ceulemans, E. (2013). CHull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behavior Research Methods*, *45*(3), 782-791.

De Raad, B., Barelds, D.P.H., Levert, E., Ostendorf, F., Mlačić, B., Di Blas, L. et al. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology*, *98,* 160–173.

De Roover, K., Ceulemans, E., & Timmerman, M. E. (2012). How to perform multiblock component analysis in practice. *Behavior Research Methods, 44,* 41−56.

De Roover, K., Ceulemans, E., Timmerman, M.E., Nezlek, J.B., & Onghena, P. (2013). Modeling differences in the dimensionality of multiblock data by means of clusterwise simultaneous component analysis. *Psychometrika, 78,* 648-668.

De Roover, K., Ceulemans, E., Timmerman, M.E., & Onghena, P. (2013). A clusterwise simultaneous component method for capturing within-cluster differences in component variances and correlations. *British Journal of Mathematical and Statistical Psychology, 86,* 81−102.

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychological Methods, 17,* 100−119.

de Winter∗, J. C. F., Dodou∗, D. I. M. I. T. R. A., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44,* 147-181.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39,* 1–38.

Diener, E., Kim-Prieto, C., Scollon, C., & Colleagues. (2001). [International College Survey 2001]. Unpublished raw data.

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling, 16,* 295–314.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria. *PeerJ PrePrints, 3,* e1350.

Eid. M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology, 81,* 869–885.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). Chichester, UK: Wiley.

Fonseca, J. R., & Cardoso, M. G. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis, 11,* 155-173.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18,* 1050–1057.

Gorsuch, R.L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research,* 25, 33-39.

Hessen, D. J., Dolan, C. V, & Wicherts, J. M. (2006). Multi-group exploratory factor analysis and the power to detect uniform bias. *Applied Psychological Research, 30,* 233–246.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 20,* 265–282.

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika, 38,* 593-604.

Jennrich, R. I., & Sampson, P. F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics, 18,* 11-17.

Jolliffe, I. T. (1986). *Principal component analysis.* New York: Springer.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Kaiser, H. F. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kiers, H. A. (1997). Techniques for rotating two or more loading matrices to optimal agreement and simple structure: A comparison and some technical details. *Psychometrika, 62,* 545-568.

Kim, E. S., Joo, S. H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement Invariance Testing Across Between-Level Latent Classes Using Multilevel Factor Mixture Modeling. *Structural Equation Modeling: A Multidisciplinary Journal.* Advance online publication.

Krysinska, K., De Roover, K., Bouwens, J., Ceulemans, E., Corveleyn, J., Dezutter, J., Duriez, B., Hutsebaut, D., & Pollefeyt, D. (2014). Measuring religious attitudes in secularized Western European context: A psychometric analysis of the post-critical belief scale. *International Journal for the Psychology of Religion, 24,* 263-281.

Kuppens, P., Ceulemans, E., Timmerman, M. E., Diener, E., & Kim-Prieto, C. (2006). Universal intracultural and intercultural dimensions of the recalled frequency of emotional experience. *Journal of Cross-Cultural Psychology, 37,* 491–515.

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician, 12,* 209–229.

Lee, S. Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 44,* 99–113.

Lorenzo-Seva, U., Kiers, H. A., & Berge, J. M. (2002). Techniques for oblique factor rotation of two or more loading matrices to a mixture of simple structure and optimal agreement. *British Journal of Mathematical and Statistical Psychology, 55,* 337-360.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods, 10,* 21.

Lukočienė, O., Varriale, R., & Vermunt, J.K. (2010). The simultaneous decision(s) about the number of lower-and higher-level classes in multilevel latent class analysis. *Sociological Methodology, 40,* 247-283.

MacKinnon, N. J., & Keating, L. J. (1989). The structure of emotions: Canada-United States comparisons. *Social Psychology Quarterly, 52,* 70–83.

Magnus, J. R., & Neudecker, H. (2007). *Matrix differential calculus with applications in statistics and econometrics* (3rd ed.). Chichester, England: Wiley & Sons.

McDonald, R. P., & Krane, W. R. (1979). A Monte Carlo study of local identifiability and degrees of freedom in the asymptotic likelihood ratio test. *British Journal of Mathematical and Statistical Psychology*, *32,* 121-132.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Notes, 17,* 1–48.

Muthén, L. K., & Muthén, B. O. (2005). *Mplus: Statistical analysis with latent variables: User's guide.* Los Angeles: Muthén & Muthén.

Ogasawara, H. (2000). Some relationships between factors and components. *Psychometrika, 65,* 167–185.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 2,* 559–572.

Rindskopf, D. (1984). Structural equation models empirical identification, Heywood cases, and related problems. *Sociological Methods & Research*, *13,* 109-119.

Rodriguez, C., & Church, A. T. (2003). The structure and personality correlates of affect in Mexico: Evidence of cross-cultural comparability using the Spanish language. *Journal of Cross Cultural Psychology, 34,* 211–230.

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

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47,* 69–76.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76,* 805–819.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Schwartz, S. H. (1994). Beyond individualism/collectivism: New cultural dimensions of values. In U. Kim, H. C. Triandis, C. Kagitcibasi, S. C. Choi, & G. Yoon (Eds.), Individualism and collectivism: Theory, methods, and applications (pp. 85–119). Thousand Oaks: Sage Publications.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Stearns, P. N. (1994). *American cool: Constructing a twentieth-century emotional style.* NYU Press.

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods, 16,* 63–79.

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

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

United Nations Development Programme. (2010). *Human Development Report 2000*. New York: Oxford University Press.

Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika, 43*, 225-243.

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research, 47*, 247–275.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological methods, 3*, 231.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*, 1–28.

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A monte carlo approach. *Multivariate Behavioral Research, 17*, 371–388.

Vermunt, J. K., & Magidson, J. (2013). Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: differential bias in representing model parameters? *Multivariate Behavioral Research, 28*, 263-311.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62*, 297–330.

Zhao, J., Jin, L., & Shi, L. (2015). Mixture model selection via hierarchical BIC. *Computational Statistics & Data Analysis, 88*, 139-153.

Zhao, J., Yu, P. L., & Shi, L. (2013). *Model selection for mixtures of factor analyzers via hierarchical BIC*. Tech. Rep, School of Statistics and Mathematics, Yunnan University of Finance and Economics, Yunnan, PR China.

Table 1. Base loading matrix and the derived cluster-specific loading matrices for Clusters 1 and 2, in case of two factors (above) and in case of four factors (below). In case of medium similarity λ_1 equals 0 and λ_2 equals 1, whereas in case of high similarity λ_1 equals √(.6) and λ_2 equals √(.4). When the number of clusters is four, the two additional loading matrices are constructed similarly, i.e., by shifting the primary loading or adding a crossloading for variables 3 and 6 for Cluster 3, and for variables 4 and 7 for Cluster 4.

	Base loading matrix		Cluster 1		Cluster 2	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Var. 1	1	0	λ_1	λ_2	1	0
Var. 2	1	0	1	0	λ_1	λ_2
Var. 3	1	0	1	0	1	0
Var. 4	1	0	1	0	1	0
Var. 5	1	0	1	0	1	0
Var. 6	1	0	1	0	1	0
Var. 7	0	1	λ_2	λ_1	0	1
Var. 8	0	1	0	1	λ_2	λ_1
Var. 9	0	1	0	1	0	1
Var. 10	0	1	0	1	0	1
Var. 11	0	1	0	1	0	1
Var. 12	0	1	0	1	0	1

	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
Var. 1	1	0	0	0	λ_1	λ_2	0	0	1	0	0	0
Var. 2	1	0	0	0	1	0	0	0	λ_1	λ_2	0	0
Var. 3	1	0	0	0	1	0	0	0	1	0	0	0
Var. 4	0	1	0	0	λ_2	λ_1	0	0	0	1	0	0
Var. 5	0	1	0	0	0	1	0	0	λ_2	λ_1	0	0
Var. 6	0	1	0	0	0	1	0	0	0	1	0	0
Var. 7	0	0	1	0	0	0	1	0	0	0	1	0
Var. 8	0	0	1	0	0	0	1	0	0	0	1	0
Var. 9	0	0	1	0	0	0	1	0	0	0	1	0
Var. 10	0	0	0	1	0	0	0	1	0	0	0	1
Var. 11	0	0	0	1	0	0	0	1	0	0	0	1
Var. 12	0	0	0	1	0	0	0	1	0	0	0	1

Table 2. Clustering of the countries of the MSFA model with two clusters and two factors per cluster for the emotion norm data from the 2001 ICS study. Except for Slovakia and South Africa, all countries are assigned to the clusters with a posterior probability $\gamma\left(z_{ik}\right)$ of 1. The probabilities for Slovakia and South Africa are given between brackets.

Cluster 1	Bangladesh, Brazil, Bulgaria, Cameroon, Georgia, Ghana, India, Iran, Nepal, Nigeria, Slovakia ($\gamma\left(z_{i1}\right)$= .9980), South Africa ($\gamma\left(z_{i1}\right)$= .9965), Thailand, Turkey, Uganda, Zimbabwe
Cluster 2	Australia, Austria, Belgium, Canada, Chile, China, Colombia, Croatia, Cyprus, Egypt, Germany, Greece, Hong Kong, Hungary, Indonesia, Italy, Japan, Kuwait, Malaysia, Mexico, Netherlands, Philippines, Poland, Portugal, Russia, Singapore, Slovenia, South Korea, Spain, Switzerland, United States, Venezuela

Table 3. Varimax (above) and echelon (below) rotated loadings of the MSFA model with two clusters and two factors per cluster for the emotion norm data from the 2001 ICS study. For each emotion, the primary loading is indicated in bold face. Below, the restricted loadings are in italic and underlined and loadings that are significantly different across clusters (according to Wald tests and after Bonferroni correction) are indicated by ** ($p < .01$) and * ($p < .05$).

Varimax rotation	Cluster 1		Cluster 2	
	Positive	Negative	Positive	Negative
Contentment	**1.44**	-0.25	**1.21**	-0.11
Happy	**1.60**	-0.26	**1.42**	-0.15
Love	**1.39**	-0.26	**1.22**	-0.06
Sad	-0.32	**1.32**	0.05	**1.26**
Jealousy (in romantic situations)	0.00	**1.29**	-0.02	**1.27**
Cheerful	**1.18**	-0.30	**1.04**	-0.05
Worry	-0.07	**1.74**	0.04	**1.43**
Stress	-0.25	**2.01**	-0.19	**1.69**
Anger	-0.37	**1.97**	-0.18	**1.54**
Pride	0.27	**1.10**	**0.60**	0.35
Guilt	0.05	**1.24**	0.11	**1.10**
Shame	0.18	**1.03**	0.08	**1.07**
Gratitude	**0.95**	-0.29	**0.86**	-0.12
Echelon rotation	Cluster 1		Cluster 2	
	Positive	Negative	Positive	Negative
Contentment	**1.44****	-0.25	**1.21****	-0.13
Happy	**1.60****	-0.26	**1.42****	-0.17
Love	**1.39***	-0.26	**1.22***	-0.08
Sad	-0.32**	**1.32**	0.07**	**1.26**
Jealousy (in romantic situations)	*0*	**1.29**	*0*	**1.27**
Cheerful	**1.18**	-0.30*	**1.04**	-0.06*
Worry	-0.07	**1.74****	0.07	**1.43****
Stress	-0.25	**2.01****	-0.16	**1.69****
Anger	-0.37	**1.97****	-0.16	**1.54****
Pride	0.27**	**1.10****	**0.61****	0.34**
Guilt	0.05	**1.24**	0.13	**1.10**
Shame	0.18	**1.03**	0.10	**1.07**
Gratitude	**0.95**	-0.29	**0.86**	-0.14

Table 4. Unique variances of the MSFA model with two clusters and two factors per cluster for the emotion norm data from the 2001 ICS study.

	Cluster 1	Cluster 2
Contentment	1.47	3.48
Happy	0.63	1.39
Love	1.21	2.37
Sad	2.76	4.19
Jealousy (in romantic situations)	2.85	4.94
Cheerful	1.53	2.38
Worry	2.01	2.86
Stress	2.15	2.63
Anger	1.87	2.23
Pride	3.41	5.33
Guilt	2.80	4.42
Shame	3.01	4.85
Gratitude	2.88	3.95

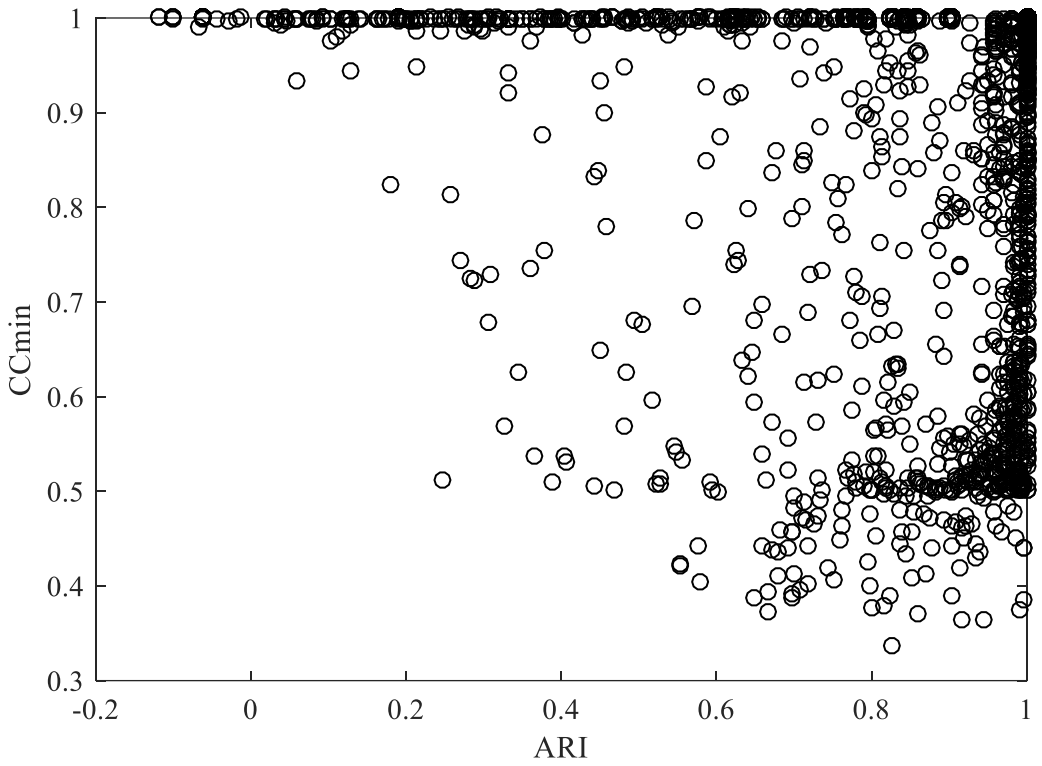
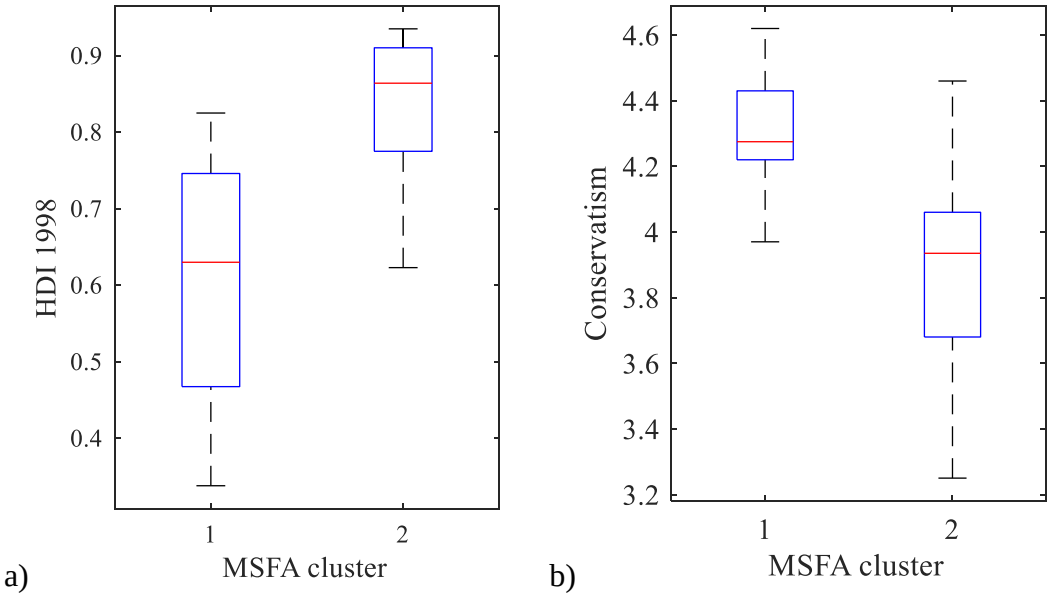


Figure 1. Scatter plot of CC_{min} versus *ARI* for the simulated data sets.



a) b)

Figure 2. Boxplots for (a) the HDI 1998 (United Nations Development Programme, 2000) and (b) the level of conservatism (Schwartz, 1994) of the countries per cluster of the MSFA model with two clusters and two factors per cluster for the ICS data set on emotion norms.

```
Cluster V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 rows
1 1 1.7473 0.7265 -0.3907 1.1339 -0.5616 -1.1998 -0.0132 -1.3467 -1.5715 -0.5305 0.6983 0.0401 80
1 2 0.1174 -0.7600 0.2907 0.4192 0.6397 0.5418 -1.2780 0.8854 -1.1675 -0.3512 0.4825 0.1812 80
1 3 -0.0833 -0.0014 2.5576 1.5717 0.9088 -1.7953 0.8549 -2.1386 -0.0093 0.2786 0.7123 0.3077 80
1 4 -0.8461 0.3073 -0.1001 -0.1164 0.2713 -1.2281 -0.9412 0.3107 -0.7565 -0.0294 1.5008 -0.5865 80
1 5 1.3313 -0.2953 0.6014 -0.0784 1.3909 -0.4921 -1.1625 -0.3562 0.5044 -1.5887 -1.4460 0.0037 80
2 6 0.1808 -0.5950 0.2237 -0.3480 0.6153 1.2128 -0.3208 -1.4779 -0.2403 -0.3889 -0.6357 1.9815 80
2 7 -2.4329 -1.7386 -2.0556 -0.4518 0.8383 -1.0102 0.0081 -0.1398 -1.0191 -0.3638 -0.2139 -0.5757 80
2 8 -1.0353 0.0475 -0.6910 0.2264 1.7928 -0.5391 -1.2320 -0.3141 0.9321 -0.5356 -1.1829 1.5041 80
2 9 -0.3471 -1.1564 -0.3303 0.3472 -0.8992 0.6269 0.0371 -0.8042 -2.3842 1.3941 -0.9446 -1.2794 80
2 10 0.3430 -0.5610 0.8318 0.4923 0.3206 0.1229 0.5291 1.3973 -0.8362 -0.1353 -1.3470 -2.3012 80
2 11 0.3889 -0.0345 -0.5115 1.2848 -2.6203 -0.1609 -1.0531 0.6125 0.6005 1.2064 0.7995 -0.2267 80
2 12 -1.2435 0.2280 -0.0794 0.6463 -0.1143 -0.5735 -0.4759 0.0054 1.4598 -0.1095 0.9843 -1.5594 80
2 13 -0.2484 0.5599 -0.0916 -1.3491 -0.1022 -1.8452 0.3866 -0.7316 0.0633 -0.7306 -0.0924 -0.6599 80
2 14 -0.3806 -1.0342 -0.4684 1.5355 -0.2831 -0.2960 -1.1959 -0.8398 1.6574 1.0114 -1.0623 -0.0474 80
2 15 2.4719 1.2264 0.3395 0.7713 -1.0924 1.7998 -1.8861 0.4930 1.4006 0.2921 -0.4422 0.2838 80
2 16 1.4124 -1.2410 -1.4496 -0.3742 0.4739 -1.5811 1.6940 1.2309 -0.2059 1.1611 -0.1257 -1.5852 80
2 17 -0.6312 -0.1502 -0.9197 -1.7419 -0.8675 -0.7788 -0.4664 -1.5448 -0.5492 -0.9583 1.0030 0.9996 80
2 18 -1.6679 0.0418 0.1085 0.1183 1.3316 -0.2125 -1.6304 -0.9432 0.1825 -1.4701 -0.2448 1.0296 80
2 19 0.6382 -1.5157 -1.3295 0.0727 -0.9773 1.1819 0.7396 -3.0663 -1.8574 -2.3692 1.6742 0.4016 80
2 20 0.9441 -0.9052 -0.4167 1.0131 -2.8936 0.3600 -1.8552 -0.3872 -0.8089 1.2214 -0.3248 -0.7658 80
```

Figure A1. ‘Example.txt’ file communicating the clustering (‘Cluster’), the number of variables (‘V2’ to ‘V13’) and the data block structure (‘V1’ and ‘rows’) to the simulation syntax for Latent Gold 5.1. Note that, in general, the number of rows may differ across data blocks.

Appendix 1: Maximum likelihood estimation of MSFA by LG 5.1.

In this appendix, we consecutively elaborate on the MSFA algorithm and the multistart procedure that we recommend to use. An example of the syntax for estimating an MSFA model in LG 5.1. is given and clarified in Appendix 2.

A1.1. Algorithm

Two of the most common algorithms for ML estimation are Expectation-Maximization (EM; Dempster, Laird, & Rubin, 1977) and Newton-Raphson (NR; Jennrich, & Sampson, 1976). In LG, a combination of both types of iterations is applied to benefit from the stability of EM when it is far from the maximum of log *L*, and the convergence speed of NR when it is close to the maximum (Vermunt & Magidson, 2013).

A1.1.1. Expectation-maximization iterations

As in all mixture models, log *L* (Equation 3) – also referred to as the 'observed-data loglikelihood' – is complicated by the latent clustering of the data blocks, making it hard to maximize log *L* directly. Therefore, the EM algorithm makes use of the so-called 'complete-data (log)likelihood', i.e., the likelihood when the cluster memberships of all data blocks are assumed to be known or, in other words, the joint distribution of the observed and latent data:

$$L(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z}) = f(\boldsymbol{X}, \mathbf{Z}; \boldsymbol{\theta}) = f(\mathbf{Z}; \boldsymbol{\theta}) f(\mathbf{X} \mid \mathbf{Z}; \boldsymbol{\theta}) \tag{7}$$

where $\mathbf{Z}$ is the $I \times K$ latent membership matrix, containing binary elements z_{ik} to indicate the cluster memberships. The probability density of the observed data conditional on the latent data is defined as follows:

$$f(\mathbf{X} \mid \mathbf{Z}; \boldsymbol{\theta}) = \prod_{i=1}^{I} \prod_{k=1}^{K} \prod_{n_i=1}^{N_i} f_k(\mathbf{x}_{n_i}; \boldsymbol{\theta}_k)^{z_{ik}} \tag{8}$$

and the probability density of the latent cluster memberships, or the so-called 'prior distribution' of the latent clustering, is given by a multinomial distribution such that:

$$f\left(\mathbf{Z};\boldsymbol{\theta}\right)=\prod_{i=1}^{I}\prod_{k=1}^{K}\pi_k^{\,z_{ik}} \tag{9}$$

with the mixing proportions π_k as the 'prior cluster probabilities'. When data block *i* belongs to cluster *k* (z_{ik} = 1), the corresponding factors in Equations 8 and 9 remain unchanged and, when the data block doesn't belong to cluster *k* (z_{ik} = 0), they become equal to one. Inserting Equations 8 and 9 into Equation 7 leads to a complete-data likelihood function containing no summation. Therefore, the complete-data loglikelihood or 'log L_c' can be elaborated as follows:

$$\begin{aligned}
\log L_c &= \log L\left(\boldsymbol{\theta}\,|\,\mathbf{X},\mathbf{Z}\right)=\log\left(\prod_{i=1}^{I}\prod_{k=1}^{K}\pi_k^{\,z_{ik}}\prod_{n_i=1}^{N_i}f_k\left(\mathbf{x}_{n_i};\boldsymbol{\theta}_k\right)^{z_{ik}}\right)\\
&=\log\left(\prod_{i=1}^{I}\prod_{k=1}^{K}\pi_k^{\,z_{ik}}\,f_k\left(\mathbf{X}_i;\boldsymbol{\theta}_k\right)^{z_{ik}}\right)\\
&=\sum_{i=1}^{I}\sum_{k=1}^{K}\left[\log\left(\pi_k^{\,z_{ik}}\right)+\sum_{n_i=1}^{N_i}z_{ik}\log\left(\frac{1}{\left(2\pi\right)^{J/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}{}'\right)\right)\right]\\
&=\sum_{i=1}^{I}\sum_{k=1}^{K}\left[z_{ik}\log\left(\pi_k\right)+z_{ik}\sum_{n_i=1}^{N_i}\left(\log\left(\frac{1}{\left(2\pi\right)^{J/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}}\right)-\frac{1}{2}\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}{}'\right)\right]\\
&=\sum_{i=1}^{I}\sum_{k=1}^{K}\left[z_{ik}\log\left(\pi_k\right)-\frac{z_{ik}}{2}\sum_{n_i=1}^{N_i}\left(J\log\left(2\pi\right)+\log\left(\left|\boldsymbol{\Sigma}_k\right|\right)+\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}{}'\right)\right]
\end{aligned} \tag{10}$$

From the summations in Equation 10, we conclude that one difficult maximization (i.e., of Equation 3) is replaced by a sequence of easier maximization problems (see Step 2 of the EM procedure). Because the values of z_{ik} are unknown, their expected values, i.e., the posterior classification probabilities $\gamma\left(z_{ik}\right)$ (Equation 2), are inserted in Equation 10, thus obtaining the expected value of log L_c or E(log L_c). Note that log L can be obtained by summing E(log L_c) over the *K* possible latent cluster assignments for each data block.

Starting from a set of initial values $\hat{\boldsymbol{\theta}}^{0}$ for the parameters, the EM procedure performs the following two steps for each iteration ν:

1. E-step: The $E(\log L_c)$-value given the current parameter estimates $\hat{\boldsymbol{\theta}}^{\nu-1}$ (i.e., $\hat{\boldsymbol{\theta}}^{0}$ when $\nu = 1$ or the estimates from the previous iteration when $\nu > 1$) is determined as follows:

   1.1. The posterior classification probabilities $\gamma(z_{ik})$ are calculated (Equation 2).

   1.2. The $\gamma(z_{ik})$-values are inserted into Equation 10 to obtain $E(\log L_c)$ for $\hat{\boldsymbol{\theta}}^{\nu-1}$.

2. M-step: The parameters $\hat{\boldsymbol{\theta}}^{\nu}$ are estimated such that $E(\log L_c)$ is maximized. Note that this also results in an increase with respect to $\log L$ (Dempster, Laird, & Rubin, 1977).

   2.1. An update of each π_k – satisfying $\sum_{k=1}^{K}\pi_k = 1$ – is given by (McLachlan & Peel, 2000):

$$\hat{\pi}_k = \frac{\sum_{i=1}^{I}\gamma(z_{ik})}{I} \tag{11}$$

   2.2. For each cluster k, the factor model for $\boldsymbol{\Sigma}_k$ is obtained by weighting each observation by the corresponding $\gamma(z_{ik})$-value and performing factor analysis on the weighted data. To this end, a separate EM algorithm (Rubin & Thayer, 1982) may be used or one of the alternatives described by Lee and Jennrich (1979). Currently, LG uses Fisher scoring to estimate the cluster-specific factor models. Fisher scoring (Lee & Jennrich, 1979) is an approximation of the NR procedure described below.

A1.1.2. Newton-Raphson iterations

In contrast to EM, NR optimization operates directly on $\log L$ (Equation 3). Specifically, NR iteratively maximizes an approximation of $\log L$, based on its first- and second-order partial derivatives, in the point corresponding to estimates $\hat{\boldsymbol{\theta}}^{\nu-1}$. Using these derivatives, NR updates

all model parameters at once. The first-order derivatives – with respect to parameters θ_r , r = 1, …, R – are gathered in the so-called 'gradient' vector **g**:

$$\boldsymbol{g} = \left[\sum_{i=1}^{I} \frac{\vartheta \log f(\mathbf{X}_i; \hat{\boldsymbol{\theta}}^{v-1})}{\vartheta \theta_1} \quad \ldots \quad \sum_{i=1}^{I} \frac{\vartheta \log f(\mathbf{X}_i; \hat{\boldsymbol{\theta}}^{v-1})}{\vartheta \theta_r} \quad \ldots \quad \sum_{i=1}^{I} \frac{\vartheta \log f(\mathbf{X}_i; \hat{\boldsymbol{\theta}}^{v-1})}{\vartheta \theta_R} \right] \tag{12}$$

where R is equal to $K - 1 + K(JQ + J)$ for MSFA with orthogonal factors. The gradient vector indicates the direction of the greatest rate of increase in log L, where element g_r is positive when higher values of log L can be found at higher values of θ_r and negative otherwise. The derivations of the elements of the gradient for an MSFA model are given in section A1.1.2.1.

The matrix of second-order derivatives – also called the 'Hessian' or **H** – contains the following elements:

$$\mathbf{H} = \left[H_{rs} \right] \text{ with } H_{rs} = \sum_{i=1}^{I} \frac{\vartheta^2 \log f(\mathbf{X}_i; \hat{\boldsymbol{\theta}}^{v-1})}{\vartheta \theta_r \vartheta \theta_s} \tag{13}$$

where H_{rs} refers to the element in row r and column s of **H**. Geometrically, the second-order derivatives refer to the curvature of the R-dimensional loglikelihood surface. Taking the curvature into account makes the update more efficient than an update based on the gradient alone (Battiti, 1992). **H** and ***g*** are combined in the NR update as follows:

$$\hat{\boldsymbol{\theta}}^{v} = \hat{\boldsymbol{\theta}}^{v-1} - \varepsilon \mathbf{H}^{-1} \mathbf{g} \tag{14}$$

where the stepsize ε, $0 < \varepsilon < 1$, is used to prevent a decrease in log L whenever a standard NR update $-\mathbf{H}^{-1}\mathbf{g}$ causes a so-called 'overshoot' (for details, see Vermunt & Magidson, 2013). The calculations of the second-order derivatives make the NR update computationally very expensive. Therefore, LG applies an approximation of the Hessian which is given in section A1.1.2.1.

A1.1.2.1. First- and second-order derivatives of observed-data loglikelihood

The first-order derivative of log *L* may be decomposed as:

$$
\begin{aligned}
&\frac{d\log L}{d\theta} = \sum_{i=1}^{I} \frac{d\log f\left(\mathbf{X}_i;\boldsymbol{\theta}\right)}{d\theta} \\
&= \sum_{i=1}^{I} \frac{1}{L_i}\frac{dL_i}{d\theta} \quad with \quad L_i = f\left(\mathbf{X}_i;\boldsymbol{\theta}\right) = \sum_{k=1}^{K} \pi_k f_k\left(\mathbf{X}_i;\boldsymbol{\theta}_k\right) = \sum_{k=1}^{K} L_{ik} \\
&= \sum_{i=1}^{I}\sum_{k=1}^{K} \frac{L_{ik}}{L_i}\frac{1}{L_{ik}}\frac{dL_{ik}}{d\theta} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{I} \gamma(z_{ik})\frac{d\log L_{ik}}{d\theta} \quad with \quad \gamma(z_{ik}) = \frac{L_{ik}}{L_i} \text{ (Equation 2)} \\
&= \sum_{k=1}^{K} \frac{d\log L_k}{d\theta}
\end{aligned}
\tag{15}
$$

where $\log L_k = \sum_{i=1}^{I} \gamma(z_{ik}) \log L_{ik}$ is the loglikelihood contribution of cluster *k*. When defining the expected observed number of blocks and number of observations in cluster *k* as $I_k = \sum_{i=1}^{I} \gamma(z_{ik})$ and $N_k = \sum_{i=1}^{I} N_i \gamma(z_{ik})$ respectively, log L_k can be expressed in terms of the cluster-specific expected observed covariance matrix $\mathbf{S}_k = \frac{1}{N_k}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i} \gamma(z_{ik}) \mathbf{x}_{n_i}{}' \mathbf{x}_{n_i}$ as follows:

$$
\begin{aligned}
&\log L_k = \sum_{i=1}^{I} \gamma(z_{ik}) \log L_{ik} = \sum_{i=1}^{I} \gamma(z_{ik}) \log\left(\pi_k f_k\left(\mathbf{X}_i;\boldsymbol{\theta}_k\right)\right) \\
&= \sum_{i=1}^{I} \gamma(z_{ik}) \left[\log\left(\pi_k\right) - \frac{1}{2}\sum_{n_i=1}^{N_i} \left(J\log\left(2\pi\right) + \log\left(\left|\boldsymbol{\Sigma}_k\right|\right) + \mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}{}' \right) \right] \\
&= I_k \log\left(\pi_k\right) - \frac{N_k}{2} J\log\left(2\pi\right) - \frac{N_k}{2}\log\left(\left|\boldsymbol{\Sigma}_k\right|\right) - \frac{1}{2}\sum_{i=1}^{I}\gamma(z_{ik})\sum_{n_i=1}^{N_i} tr\left(\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_{n_i}{}'\right) \\
&= I_k \log\left(\pi_k\right) - \frac{N_k}{2} J\log\left(2\pi\right) - \frac{N_k}{2}\log\left(\left|\boldsymbol{\Sigma}_k\right|\right) - \frac{1}{2} tr\left(\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\gamma(z_{ik})\mathbf{x}_{n_i}{}'\mathbf{x}_{n_i}\boldsymbol{\Sigma}_k^{-1}\right) \\
&= I_k \log\left(\pi_k\right) - \frac{N_k}{2}\left(J\log\left(2\pi\right) + \log\left(\left|\boldsymbol{\Sigma}_k\right|\right) + tr\left(\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1}\right)\right)
\end{aligned}
\tag{16}
$$

The first derivative of $\log L_k$ thus becomes the following (Magnus & Neudecker, 2007):

$$
\begin{aligned}
\frac{d\log L_k}{d\theta} &= I_k \frac{d\log(\pi_k)}{d\theta} - \frac{N_k}{2}\left(\frac{d\log(|\boldsymbol{\Sigma}_k|)}{d\theta} + tr\left(\frac{d\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1}}{d\theta}\right)\right) \\
&= \frac{I_k}{\pi_k}\frac{d\pi_k}{d\theta} - \frac{N_k}{2}\left(tr\left(\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right) + tr\left(\frac{d\mathbf{S}_k}{d\theta}\boldsymbol{\Sigma}_k^{-1} + \mathbf{S}_k\frac{d\boldsymbol{\Sigma}_k^{-1}}{d\theta}\right)\right) \\
&= \frac{I_k}{\pi_k}\frac{d\pi_k}{d\theta} - \frac{N_k}{2}\left(tr\left(\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right) + tr\left(-\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\boldsymbol{\Sigma}_k^{-1}\right)\right) \quad with \quad \frac{d\mathbf{S}_k}{d\theta} = 0 \quad (observed) \\
&= \frac{I_k}{\pi_k}\frac{d\pi_k}{d\theta} + \frac{N_k}{2}\left(tr\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right) - tr\left(\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)\right) \\
&= \frac{I_k}{\pi_k}\frac{d\pi_k}{d\theta} + \frac{N_k}{2}\left(tr\left(\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\right)\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)\right) \\
&= \frac{I_k}{\pi_k}\frac{d\pi_k}{d\theta} + \frac{N_k}{2}\left(vec\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\right)' vec\left(\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)\right),
\end{aligned}
\tag{17}
$$

such that $\frac{d\log L}{d\theta} = \sum_{k=1}^{K}\frac{I_k}{\pi_k}\frac{d\pi_k}{d\theta} + \sum_{k=1}^{K}\frac{N_k}{2}\left(vec\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\right)' vec\left(\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)\right)$. The second-order derivative of $\log L_k$ is then equal to (Magnus & Neudecker, 2007):

$$
\begin{aligned}
\frac{d^2\log L_k}{d\theta d\theta'} &= \frac{N_k}{2}\left(tr\left(\frac{d}{d\theta'}\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\right)\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)\right) \\
&= \frac{N_k}{2}tr\left(\frac{d}{d\theta'}\left(\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)\right) \\
&= \frac{N_k}{2}tr\left(\begin{array}{l}
\frac{d\boldsymbol{\Sigma}_k^{-1}}{d\theta'}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta} + \boldsymbol{\Sigma}_k^{-1}\frac{d}{d\theta'}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta} \\
+\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\frac{d\boldsymbol{\Sigma}_k^{-1}}{d\theta'}\frac{d\boldsymbol{\Sigma}_k}{d\theta} + \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\frac{d}{d\theta'}\left(\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)
\end{array}\right) \\
&= \frac{N_k}{2}tr\left(\begin{array}{l}
\frac{d\boldsymbol{\Sigma}_k^{-1}}{d\theta'}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta} + \boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta'}\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta} \\
+\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\frac{d\boldsymbol{\Sigma}_k^{-1}}{d\theta'}\frac{d\boldsymbol{\Sigma}_k}{d\theta} + \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\frac{d}{d\theta'}\left(\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)
\end{array}\right).
\end{aligned}
\tag{18}
$$

Because the expected value of $\left(\mathbf{S}_k - \boldsymbol{\Sigma}_k\right)$ equals zero, the expected value of the second derivative of $\log L_k$ becomes $E\left(\frac{d^2\log L_k}{d\theta d\theta'}\right) = \frac{N_k}{2}tr\left(\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta'}\boldsymbol{\Sigma}_k^{-1}\frac{d\boldsymbol{\Sigma}_k}{d\theta}\right)$. Therefore, within LG, the second-order derivative of $\log L$ is approximated as:

$$\frac{d^2 \log L}{d\theta d\theta'} = \sum_{k=1}^{K} \frac{d^2 \log L_k}{d\theta d\theta'} = \sum_{k=1}^{K} \frac{N_k}{2} tr\left(\boldsymbol{\Sigma}_k^{-1} \frac{d\boldsymbol{\Sigma}_k}{d\theta'} \boldsymbol{\Sigma}_k^{-1} \frac{d\boldsymbol{\Sigma}_k}{d\theta} \right). \quad (19)$$

A1.1.3. Convergence

In practice, the estimation process starts with a number of EM iterations. When close to the final solution, the program switches to NR iterations to speed up convergence. Convergence can be evaluated with respect to log *L* or with respect to the parameter estimates. LG applies the latter approach (Vermunt & Magidson, 2013). More specifically, convergence is evaluated by computing the following quantity:

$$\delta = \sum_{r=1}^{R} \left| \frac{\hat{\theta}_r^v - \hat{\theta}_r^{v-1}}{\hat{\theta}_r^{v-1}} \right|, \quad (20)$$

which is the sum of the absolute value of the relative changes in the parameters. The convergence criterion that is used for MSFA in this paper is $\delta < 1\times10^{-8}$. The iteration also stops when the change in log *L* is negligible, i.e., smaller than 1×10^{-12}.

It is important to note that, when convergence is reached, this is not necessarily a global maximum. To increase the probability of finding the global maximum, a multistart procedure is used, which is described in the next section.

A1.2. Multistart procedure

LG applies a tiered testing strategy with respect to sets of starting values (Vermunt & Magidson, 2013). Specifically, it starts from a user-specified number of sets (say 25), each of which is updated for a maximum number of iterations (e.g., 100) or until δ is smaller than a specified criterion (e.g., 1×10^{-5}). Subsequently, it continues with the 10% (rounded upwards)

most promising sets (i.e., with the highest log *L*), performing another two times the specified number of iterations (e.g., 2 × 100). Finally, it continues with the best solution until convergence. Note that such a procedure increases considerably the probability of finding the global ML solution, but does not guarantee it. Thus, one should remain cautious of local maxima.

With respect to generating sets of starting values, a special option was added to the LG 5.1 syntax module to create suitable initial values for the cluster-specific loadings and unique variances of MSFA. Specifically, the initial values are based on the loadings and residual variances of a principal component (PCA) model (Jolliffe, 1986; Pearson, 1901), in principal axes position, for the entire data set. This seems reasonable as typically loadings from PCA strongly resemble the ones of EFA (Widaman, 1993). To create *K* sufficiently different sets of initial factor loadings, randomness is added to the PCA loadings for each cluster *k*:

$$\mathbf{\Lambda}_k = (.25 + rand(1)) * \mathbf{\Lambda}_{PCA} \quad for\ k = 1, ..., K \tag{21}$$

where '*rand*(1)' indicates a $J \times Q$ matrix of random numbers sampled from a uniform distribution between 0 and 1, and '*' denotes the elementwise product. Note that the default random seed is based on time, such that the added random numbers will be unique for each set of starting values (Vermunt & Magidson, 2013). To avoid the occurrence of Heywood cases (Rindskopf, 1984; Van Driel, 1978) very early in the model estimation, the initial unique variances are generated as follows:

$$diag(\mathbf{D}_k) = \text{var}(\mathbf{E}_{PCA}) * 1.5 \quad for\ k = 1, ..., K, \tag{22}$$

where diag($\mathbf{D}_k$) refers to the diagonal elements of $\mathbf{D}_k$ and $\mathbf{E}_{PCA}$ is the matrix of PCA residuals. Preliminary simulation studies indicated a much lower sensitivity to local maxima and a faster computation time when using these starting values instead of mere random values.

Appendix 2: Latent Gold 5.1 syntax for MSFA analysis

```
options
   algorithm
      tolerance=1e-008 emtolerance=1e-006 emiterations=2500
      nriterations=500;
   startvalues
      seed=0 sets=25 tolerance=1e-005 iterations=100 PCA;
   bayes
      categorical=0 variances=0 latent=0 poisson=0;
   montecarlo
      seed=0 replicates=500 tolerance=1e-008;
   quadrature  nodes=10;
   missing  excludeall;
   output
      iterationdetail classification parameters=effect standarderrors
      probmeans=posterior profile bivariateresiduals
      writeparameters='results_parameters.csv' write='results.csv'
      writeloadings='results_loadings.txt';
   outfile 'classif.txt' classification;
variables
   groupid V1;
   dependent V2 continuous, V3 continuous, V4 continuous, V5 continuous, V6
      continuous, V7 continuous, V8 continuous, V9 continuous, V10
      continuous, V11 continuous, V12 continuous, V13 continuous;
   independent V1 nominal;
   latent
      F1 continuous,
      F2 continuous,
      F3 continuous,
      F4 continuous,
      Cluster nominal group 2 coding=first;
equations
// factor variances
   (1) F1;
   (1) F2;
   (1) F3;
   (1) F4;
// logistic regression model for clusters: contains only intercept
   Cluster <- 1;
// regression models for items: no intercept and loading which vary across
clusters
   V2 - V13 <- F1 | Cluster + F2 | Cluster + F3 | Cluster + F4 | Cluster;
// item variances
   V2 - V13 | Cluster;
```

The LG syntax is built up from three sections, i.e., ‘options, ‘variables’, and ‘equations’. Firstly, the ‘options’ section pertains to specifications regarding the estimation process and to output options. The parameters in the ‘algorithm’ subsection indicate when the algorithm should proceed with NR instead of EM iterations and when convergence is reached (see Vermunt & Magidson, 2013). The ‘startvalues’ subsection includes the parameters pertaining to the multistart procedure (Section A1.2). Specifically, for each set of starting values (the

number of sets is specified by ‘sets’), the model is re-estimated for as many iterations as specified by ‘iterations’ or until δ drops below the ‘tolerance’ value. Then, the multistart procedure proceeds as described in Section A1.2. ‘PCA’ prompts LG to use the PCA-based starting values, whereas otherwise ‘seed=0’ would give the default random starts (for other possible ‘seed’ values, see Vermunt & Magidson, 2013). In the ‘output’ and ‘outfile’ subsections, the desired output can be specified by the user (for more details, see Vermunt & Magidson, 2013). The parameters of the remaining subsections are not used in this paper.

Secondly, the ‘variables’ section specifies the different types of variables included in the model. Since MSFA operates on multilevel data, after ‘groupid’, the variable in the data file that specifies the group structure (i.e., the data block number for each observation) should be specified (i.e., ‘V1’), using its label in the data file. In the ‘dependent’ subsection, the dependent variables of the model (i.e., the observed variables) should be specified, by means of their label in the data file and their measurement scale. Next, the ‘independent’ variables can be specified. In the MSFA case, it is useful to include the grouping variable as an independent variable, in order to get the cluster memberships per data block rather than per row (i.e., in the ‘probmeans-posterior’ output tab; Vermunt & Magidson, 2013). Finally, the ‘latent’ variables of the MSFA model are the factors (i.e., ‘F1’ to ‘F4’ in the example syntax) and the mixture model clustering (i.e., ‘Cluster’). In particular, the former are specified as continuous latent variables, whereas the latter is specified as a nominal latent variable at the group level with a specified number of categories (i.e., the desired number of clusters). By ‘coding=first’ cluster 1 is (optionally) coded as the reference cluster in the logistic regression model for the clustering (explained below). For other coding options, see Vermunt and Magidson (2013).

In the ‘equations’ section, the model equations are listed. First, the factor variances are specified and fixed at one. No factor covariances are specified, implying that orthogonal factors are estimated. Note that both restrictions apply to each data block, because we do not specify

an effect of the grouping variable on the factor (co)variances. Next, a logistic regression model for the categorical latent variable ‘Cluster’ is specified (Vermunt & Magidson, 2013), which contains only an intercept term in case of MSFA. Specifically, this intercept vector relates to the prior probabilities or mixing proportions of the clusters in that it includes the odds ratio’s for the $K-1$ non-reference clusters with respect to the reference cluster, i.e., cluster 1:

$$odds_k = \log\left(\frac{\pi_k}{\pi_1}\right). \tag{23}$$

Then, regression models are defined for the observed variables, i.e., which variables are regressed on which factors. Note that, for MSFA, all variables are regressed on all factors (i.e., it applies EFA) and that no intercept term is included. By default, overall factor means are equal to zero and no effect is specified to make them differ between data blocks or clusters. To obtain factor loadings that differ between the clusters, ‘| Cluster’ is added to each regression effect. Finally, item variances are added, which pertain to the unique variances in this case and which are also allowed to differ across clusters. Optionally, at the end of the syntax, additional restrictions may be specified or starting values for all parameters may be given, either by directly typing them in the syntax or by referring to a text file (see Vermunt & Magidson, 2013).

Appendix 3: Latent Gold 5.1 syntax for MSFA simulation

```
//LG5.1//
version = 5.1
infile 'C:\Users\Documents\...\example.txt' quote = single

model
options
   algorithm
      tolerance=1e-008 emtolerance=1e-008 emiterations=2500 nriterations=0;
   startvalues
      seed=0 sets=1 tolerance=1e-005 iterations=50;
   bayes
      categorical=1 variances=1 latent=1 poisson=1;
   montecarlo
      seed=0 replicates=500 tolerance=1e-008;
   quadrature  nodes=10;
   missing  excludeall;
   output
      iterationdetail classification parameters=effect standarderrors
      probmeans=posterior profile bivariateresiduals;
   outfile 'simulateddata.txt' simulation;
variables
   caseweight rows;
   groupid V1;
   dependent V2 continuous, V3 continuous, V4 continuous, V5 continuous, V6
      continuous, V7 continuous, V8 continuous, V9 continuous, V10
      continuous, V11 continuous, V12 continuous, V13 continuous;
   independent V1 nominal;
   independent Cluster nominal;
   latent
      F1 continuous,
      F2 continuous,
      F3 continuous,
      F4 continuous;
equations
// factor variances
   (1) F1;
   (1) F2;
   (1) F3;
   (1) F4;
// regression models for items: no intercept and loading which vary across
clusters
   V2 - V13 <- F1 | Cluster + F2 | Cluster + F3 | Cluster + F4 | Cluster;
// item variances
   V2 - V13 | Cluster;
// starting values
   'startingvalues.txt'
end model
```

For generating the simulated data sets by means of LG, syntaxes were used like the one shown above. The cluster memberships, the data block sizes (i.e., the number of rows per block, factor 2), as well as the number of variables (including a variable to identify the data blocks) were communicated to the simulation syntax by means of a text file (Figure A1) which is

referred to as the ‘example’ file in the LG manual (Vermunt & Magidson, 2013). The observed variables are still to be simulated and can thus take on arbitrary but admissible values in the example file; in the current simulation study, random numbers from a standard normal distribution were used. The simulation syntax lists a lot of technical parameters in the ‘Options’ section. Most of them are discussed in Appendix 2. The ‘outfile simulateddata.txt simulation’ option will generate one simulated data set from the population model that is specified further on in the syntax, and will save it as a text file. The montecarlo parameters pertain to other types of simulation studies and resampling studies (see Vermunt & Magidson, 2013). The MSFA population model encompasses a model syntax (see Appendix 2) and ‘starting values’ for all free model parameters (i.e., the population-level parameter values that were written into a text file, with, per cluster, first the unique variances and then the loadings of the first factor, followed by the loadings of the second factor, and so on; see Figure A1). The model syntax determines the data block structure of the data to be simulated by the ‘groupid’ and ‘caseweight’ variable. An important difference with an analysis is that, when simulating, the clustering is known (through the example file) and it is thus defined as an independent variable in the simulation syntax model.

[Insert Figure A1 about here]