British Journal of Mathematical and Statistical Psychology (2009), 62, 299–318
© 2009 The British Psychological Society



The British Psychological Society

www.bpsjournals.co.uk

Bootstrap confidence intervals in multi-level simultaneous component analysis

Marieke E. Timmerman[1]*, Henk A. L. Kiers[1], Age K. Smilde[2], Eva Ceulemans[3] and Jeroen Stouten[3]

[1]Heymans Institute for Psychology, University of Groningen, Groningen, The Netherlands

[2]Biosystems Data Analysis, University of Amsterdam, Amsterdam, The Netherlands

[3]University of Leuven, Leuven, Belgium

Multi-level simultaneous component analysis (MLSCA) was designed for the exploratory analysis of hierarchically ordered data. MLSCA specifies a component model for each level in the data, where appropriate constraints express possible similarities between groups of objects at a certain level, yielding four MLSCA variants. The present paper discusses different bootstrap strategies for estimating confidence intervals (CIs) on the individual parameters. In selecting a proper strategy, the main issues to address are the resampling scheme and the non-uniqueness of the parameters. The resampling scheme depends on which level(s) in the hierarchy are considered random, and which fixed. The degree of non-uniqueness depends on the MLSCA variant, and, in two variants, the extent to which the user exploits the transformational freedom. A comparative simulation study examines the quality of bootstrap CIs of different MLSCA parameters. Generally, the quality of bootstrap CIs appears to be good, provided the sample sizes are sufficient at each level that is considered to be random. The latter implies that if more than a single level is considered random, the total number of observations necessary to obtain reliable inferential information increases dramatically. An empirical example illustrates the use of bootstrap CIs in MLSCA.

1. Introduction

Multi-level simultaneous component analysis (MLSCA; Timmerman, 2006) is an exploratory component analysis of hierarchically ordered data. Such data occur when, for example, the individuals observed are nested within groups. MLSCA specifies a component model for each level in the data, like the group and the individual levels, and examines the linear relationships between variables at each level. Similarities between

* *Correspondence should be addressed to Dr Marieke E. Timmerman, Heymans Institute for Psychology (DPMG), 9712 TS Groningen, The Netherlands (e-mail: m.e.timmerman@rug.nl).*

DOI:10.1348/000711007X265894

groups of objects at a level can be expressed by applying certain constraints in the component model, using the approach of simultaneous component analysis (SCA; Kiers & Ten Berge, 1994; Timmerman & Kiers, 2003). MLSCA has proved to be useful in a range of applications, such as trait/state personality psychology (Timmerman, 2006), cross-cultural psychology (Kuppens, Ceulemans, Timmerman, Diener, & Kim-Prieto, 2006), functional genomics (Jansen, Hoefsloot, Van der Greef, Timmerman, & Smilde, 2005), and process control (De Noord & Theobald, 2005).

If an MLSCA is performed on sample data, it is important to assess the inferential properties of the model estimates. A method for assessing the overall model fit using a cross-validation procedure has been described (Timmerman, 2006). However, inferential information on the individual parameter estimates is lacking thus far. This paper aims to fiu this gap. In particular, we use the bootstrap methodology (Davison & Hinkley, 1997; Efron & Tibshirani, 1993) for estimating confidence intervals (CIs) for the individual parameters.

The bootstrap is a resampling-based approach for estimating the uncertainty of parameter estimates. Because various implementations of the bootstrap principles are possible, many different bootstrap procedures can be defined. As different procedures generally yield different results, it is important to carefully select a specific procedure that harmonizes with the nature of the data. In this paper, we discuss considerations for selecting a proper MLSCA bootstrap procedure. The two main issues to address are the type of resampling scheme and the non-uniqueness of the parameters. The quality of resulting bootstrap CIs in finite samples is examined in a simulation study, and the result of an empirical example is shown.

2. Multi-level simultaneous component analysis

Multi-level simultaneous component analysis results in a component model for each level in hierarchically ordered data. A two-level example is when pupils are nested within classes. The pupils are said to be at level 1, and the classes at level 2. MLSCA is of use for studying the linear relationships between the variables observed, both within and between the classes. The interest in relationships between all variables contrasts MLSCA with multi-level regression (see Bryk & Raudenbush, 1992; Snijders & Bosker, 1999), in which the predictive value of observed explanatory variables for a dependent variable is examined.

MLSCA bears more resemblance to multi-group and multi-level structural equation models (SEMs; see Du Toit & Du Toit, 2008; Jöreskog, 1971; Muthén, 1989). The key difference between SEMs and component analysis (CA), of which MLSCA is a variant, is the definition of the factors. In CA, the factors (components) are based on the observed variables, and in SEMs on the common parts of the variables. Thus, in CA, the distinction between common and unique parts of variables is simply ignored. Furthermore, the SEMs and CA traditions differ, for instance in the typical estimation procedures (maximum likelihood vs. least squares), distributional assumptions (strong vs. weak), and modelling approach (confirmatory vs. exploratory), but those approaches are by no means intrinsic to either SEMs or CA.

Although the explicit modelling of unique variance in SEMs results in a more complete model, an empirical SEM is not necessarily better than a CA model. Because the number of free parameters is limited, all empirical models inevitably suffer from model error. In the SEM/CA context, model error is the difference between the

population covariance matrix and the model implied covariance matrix (Browne & Cudeck, 1992). A good model explains the main phenomena under study, in accordance with the (acknowledged) theory about the subject under study. In practice, a CA model or SEM may serve the same purpose, and the approaches can be used complementarily. However, sample size limitations, problematic data sizes (e.g. very large number of variables), or severe violations of distributional assumptions may hinder the successful use of SEMs. Because a CA solution can always be obtained, CA may be the only option to arrive at an insightful solution in some cases. This section introduces the MLSCA models and provides an empirical example to illustrate the use of MLSCA.

2.1. MLSCA models

For ease of presentation, the MLSCA models (Timmerman, 2006) for two levels will be presented. The principles used here can be applied straightforwardly to serve more than two levels. The level 2 units will be denoted by 'groups' and level 1 units by 'individuals' in what follows. MLSCA for such two-level data results in separate component models for the between and within parts. The between part model covers differences between the groups, and the within part model the differences between individuals within groups. Let us collect the observed scores y_{ijk_i} of individual k_i in group i ($k_i = 1, \ldots, K_i$; $i = 1, \ldots, I$) on variable j ($j = 1, \ldots, J$) in data matrix $\mathbf{Y}_i$ ($K_i \times J$). In MLSCA, the data matrix $\mathbf{Y}_i$ is decomposed as

$$\mathbf{Y}_i = \mathbf{1}_{K_i}\mathbf{m}' + \mathbf{1}_{K_i}\mathbf{f}'_{ib}\mathbf{B}'_b + \mathbf{F}_{iw}\mathbf{B}'_w + \mathbf{E}_i, \tag{1}$$

where $\mathbf{1}_{K_i}$ is the $K_i \times 1$ vector with each element equal to 1, $\mathbf{m}$ ($J \times 1$) is the vector containing the offsets of the J variables across all individuals in all groups, $\mathbf{f}_{ib}$ ($Q_b \times 1$) denotes the vector with between component scores of group i, $\mathbf{B}_b$ ($J \times Q_b$) denotes the between loading matrix, $\mathbf{F}_{iw}$ ($K_i \times Q_w$) denotes the within component scores matrix of group i of individuals $1, \ldots, K_i$, $\mathbf{B}_w$ ($J \times Q_w$) denotes the within loading matrix, and $\mathbf{E}_i$ ($K_i \times J$) denotes the matrix of residuals; Q_b is the number of between components and Q_w the number of within components. For identification purposes it is required that $\sum_{i=1}^{I} K_i \mathbf{f}_{ib} = \mathbf{0}_{Q_b}$ and $\mathbf{1}'_{K_i}\mathbf{F}_{iw} = \mathbf{0}'_{Q_w}$, $i = 1, \ldots, I$. The latter constraints are sufficient to ensure that the offset term, the between part and the within part of the model are uniquely separated. The between component scores $\mathbf{f}'_{ib}$, $i = 1, \ldots, I$, can be conveniently arranged in the matrix $\mathbf{F}_b$ ($I \times Q_b$).

Because the within loading matrix $\mathbf{B}_w$ is the same for all I groups, the bases on which the within component scores (in $\mathbf{F}_{iw}$) are expressed are equal for all I groups, giving the same interpretation of the within components for all groups. To partly identify the model, and without loss of generality, the mean covariance matrix of the within component scores matrices ($\mathbf{F}_{iw}$, $i = 1, \ldots, I$) across the I groups is fixed at $\mathbf{I}_{Q_w}$, the ($Q_w \times Q_w$) identity matrix. The covariance matrix of the within component score matrix $\mathbf{F}_{iw}$ offers information on the relative degree of similarity of individuals within group i on the within components, and on the degree of linear relationship(s) between within components. Possible similarities in the covariance matrices of within component scores over the I groups are modelled by imposing particular constraints (Timmerman, 2006; Timmerman & Kiers, 2003). Thus, four MLSCA variants have been defined. Ordered from the most to least constrained, they are: MLSCA-ECP, with both covariances and variances equal across groups; MLSCA-IND, with zero covariances and free variances; MLSCA-PF2, with equal covariances and free variances; and MLSCA-P, with free covariance matrices.

To fit the models within the MLSCA framework, the sum of squared residuals is minimized. As proved in Timmerman (2006), globally optimal estimates can be obtained in two steps. The 'analysis of variance (ANOVA) step' decomposes the observed data into the overall mean, a between part and a within part. The 'component analysis step' performs CA variants on both the between and the within part. The resulting model estimates are non-unique to different extents. In all MLSCA variants, the between loading matrix $\mathbf{B}_b$ has transformational freedom. This implies that it may be orthogonally or obliquely transformed without changing the explained sum of squares, provided that such a transformation is compensated for in the between component scores $\mathbf{f}_{ib}$, $i = 1, \ldots, I$. The non-uniqueness of the estimates of the within part depends on the particular MLSCA variant. That is, the within loading matrices of MLSCA-P and MLSCA-ECP both have transformational freedom. The within loading matrices of MLSCA-PF2 and MLSCA-IND are essentially unique in practice, meaning that the estimates are unique up to permutation and reflection (see Timmerman & Kiers, 2003, for uniqueness conditions).

2.2. Empirical example of MLSCA

To illustrate MLSCA, we present an empirical example from social dilemma research. In this experimental study, 283 participants played a public good dilemma game with an opponent. In the game, both players received a bonus. The experimental design consisted of two fully crossed factors. Firstly, the amount of the bonus was either equal for both players (equal condition) or higher for the opponent (unequal condition). Secondly, some of the participants received a message about a negative personal event that happened to the opponent, and they were asked to imagine how the opponent felt (high empathy condition); some of the participants also received the negative personal event message but were asked to take an objective perspective (low empathy conditions); some of the participants did not receive a message (control conditions). The 283 participants were assigned randomly to one of the six resulting equality/empathy conditions. After playing the dilemma game, participants were asked how they felt using 21 emotion terms, to be rated on a 7-point Likert scale. This study resulted in a 283 participants by 21 emotions data matrix, where each of the participants (level 1) belonged to a particular equality/empathy condition (level 2).

To gain insight into how the affective reactions of the participants were determined by the manipulated dilemma features on the one hand, and individual differences in, for instance, personality on the other hand, MLSCA analyses were applied. On the basis of the between and within scree plots in Figure 1 and interpretability, an MLSCA-ECP solution was selected with two between components and two within components. This solution accounts for 98.7% of the between variance in emotional experience and 42.2% of the within variance, which respectively correspond to 53.7% and 19.2% of the total variance in the data. The between and within loadings, both in principal axes position, appeared to be well interpretable. Inferential information on the between component scores, the between loadings and the within loadings was obtained by estimating 95% CIs. To this end, the bias-corrected and accelerated (BCa) bootstrap procedure with stratified resampling (using 1,000 bootstrap replicates) and orthogonal Procrustes rotation of the bootstrap loadings towards the sample loadings was used; the details of this procedure will be discussed in detail later.

The between part of the selected MLSCA-ECP solution summarizes the differences in the average emotional profiles of the participants in the six equality/empathy

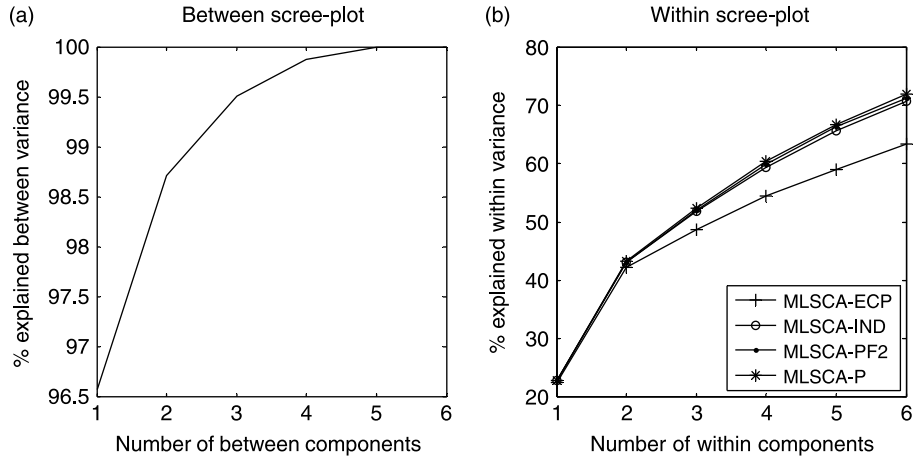

Figure 1. Percentages of (a) explained between variance for MLSCA solutions with the numbers of between components varying from 1 to 6 and (b) explained within variance for MLSCA-ECP, MLSCA-IND, MLSCA-PF2, and MLSCA-P solutions with the numbers of within components varying from 1 to 6.

conditions. Given the between loadings and corresponding 95% CIs in Table 1, the two between components can be labelled 'positive versus negative affect' and 'empathy', respectively. The component scores of the six conditions on these between components and the associated 95% CIs, as presented in Figure 2, yield the following insight into the effects of the experimental manipulations. Firstly, there is a main effect of the equality manipulation on the valence of the experienced emotions, where equal conditions elicited more positive affective reactions than unequal conditions. Neither the main effect of the empathy manipulation nor an interaction effect between equality and empathy appears to be present for positive versus negative affect. Secondly, there is a main effect of the empathy manipulation on whether or not participants feel empathy: The average empathic component score of the high empathy conditions is higher than

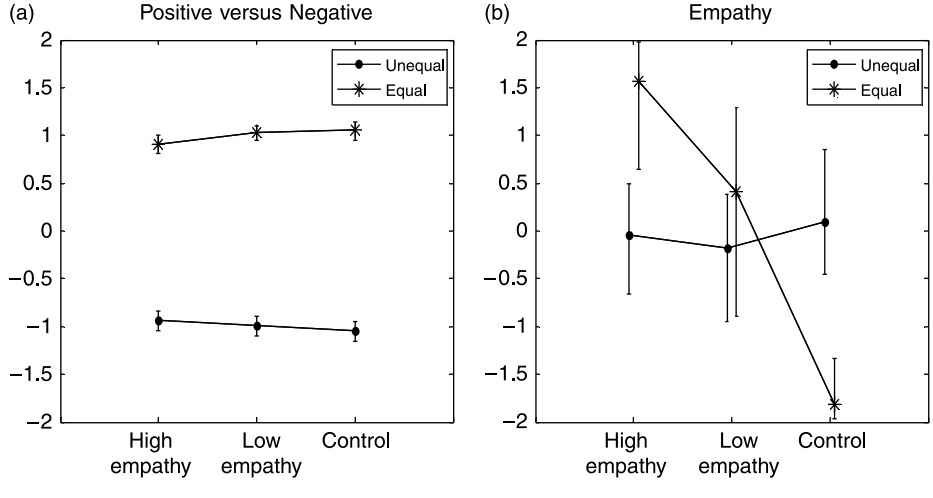

Figure 2. Component scores of the conditions on the two between components, with error bars indicating 95% BCa bootstrap CIs.

that of the control conditions, with the average score of the low empathy conditions being in between. However, this main effect is qualified by the presence of an interaction effect of the equality and empathy manipulations: Whereas the component scores in the equal conditions follow the pattern just described, the empathy manipulation has no clear effect on the experience of empathic emotions in the unequal conditions.

The within part of the MLSCA-ECP solution gives us summary information about how the individual emotional profiles of the participants differ within a condition. The within loadings and associated 95% CIs are presented in Table 1. When interpreting these within loadings and comparing them to the between loadings, one should take into account that the squared between and within loadings indicate the proportion of the total variance of the particular individual emotion that is explained by the component at hand. As such, it can be concluded from Table 1 that the first within component can again be labelled positive versus negative affect, but that this component explains a considerably smaller part of the total variance than its between counterpart. With respect to the second within component, it is clear that this component is characterized by high loadings for empathic emotions such as 'compassionate' and 'concern' as well as positive interpersonal emotions such as 'tender-hearted' and 'tender'; we label this within component 'sympathy'. Importantly, the differences in the amount of sympathy are to a larger extent determined by characteristics of the individual participants than the differences in the valence of the affective reactions, as can be seen from the clear

Table 1. Loadings of the emotion terms on the between and within components, with 95% BCa bootstrap CIs in brackets. Loadings significantly deviating from 0 ($\alpha = .05$) are printed in italics

	Between components		Within components	
Emotion terms	Positive versus negative affect	Empathy	Positive versus negative affect	Sympathy
Angry	*−0.88* [−0.92, −0.84]	−0.03 [−0.06, 0.00]	*−0.32* [−0.40, −0.23]	0.00 [−0.05, 0.05]
Hurt	*−0.79* [−0.84, −0.73]	−0.04 [−0.07, 0.01]	*−0.32* [−0.41, −0.24]	0.08 [−0.00, 0.15]
Irritated	*−0.89* [−0.93, −0.86]	−0.03 [−0.07, 0.01]	*−0.29* [−0.36, −0.20]	0.03 [−0.01, 0.09]
Sympathetic	*0.22* [0.03, 0.51]	−0.07 [−0.21, 0.05]	*0.27* [0.17, 0.37]	*0.37* [0.29, 0.48]
Elated	*0.82* [0.77, 0.87]	−0.02 [−0.08, 0.04]	*0.29* [0.21, 0.36]	*0.16* [0.09, 0.23]
Warm	*0.79* [0.74, 0.84]	−0.01 [−0.09, 0.06]	*0.32* [0.25, 0.40]	*0.18* [0.10, 0.27]
Annoyed	*−0.88* [−0.92, −0.84]	0.00 [−0.03, 0.05]	*−0.37* [−0.44, −0.28]	−0.01 [−0.06, 0.06]
Compassionate	−0.03 [−0.14, 0.08]	*0.36* [0.26, 0.47]	−0.11 [−0.24, 0.01]	*0.64* [0.58, 0.72]
Fearful	*−0.39* [−0.46, −0.29]	0.00 [−0.09, 0.10]	*−0.35* [−0.47, −0.24]	*0.44* [0.34, 0.52]
Frustrated	*−0.91* [−0.94, −0.87]	−0.03 [−0.07, −0.00]	*−0.30* [−0.38, −0.22]	0.01 [−0.02, 0.07]
Surprised	*−0.81* [−0.87, −0.77]	−0.05 [−0.14, 0.02]	*−0.18* [−0.27, −0.10]	*0.12* [0.03, 0.20]
Tender-hearted	−0.04 [−0.14, 0.09]	0.12 [−0.09, 0.31]	−0.08 [−0.20, 0.04]	*0.76* [0.69, 0.83]
Satisfied	*0.91* [0.87, 0.93]	−0.02 [−0.06, 0.01]	*0.30* [0.24, 0.37]	*0.06* [0.02, 0.11]
Disappointed	*−0.89* [−0.93, −0.85]	−0.02 [−0.05, 0.01]	*−0.32* [−0.40, −0.23]	0.00 [−0.04, 0.06]
Relieved	*0.66* [0.59, 0.73]	−0.05 [−0.16, 0.05]	*0.36* [0.26, 0.43]	*0.31* [0.21, 0.41]
Indignant	*−0.90* [−0.93, −0.86]	0.01 [−0.03, 0.05]	*−0.31* [−0.38, −0.23]	−0.01 [−0.06, 0.05]
Happy	*0.89* [0.86, 0.92]	−0.02 [−0.06, 0.02]	*0.32* [0.25, 0.39]	*0.06* [0.01, 0.11]
Concerned	*−0.25* [−0.35, −0.15]	*0.27* [0.14, 0.37]	*−0.24* [−0.34, −0.13]	*0.63* [0.57, 0.71]
Enraged	*−0.78* [−0.83, −0.74]	−0.04 [−0.08, 0.00]	*−0.39* [−0.47, −0.32]	0.00 [−0.05, 0.07]
Tender	*0.46* [0.37, 0.56]	−0.09 [−0.29, 0.14]	*0.28* [0.19, 0.37]	*0.50* [0.40, 0.58]
Hostile	*−0.75* [−0.80, −0.70]	−0.05 [−0.09, 0.01]	*−0.37* [−0.47, −0.28]	0.01 [−0.07, 0.10]

differences in sizes of the within loadings of the emotions characterizing sympathy and positive versus negative affect. Regarding the within component scores, the selection of the MLSCA-ECP solution implies that the (co)variances of these scores are equal across the different experimental conditions. Eliminating this equality constraint, by performing an MLSCA-IND analysis, yielded approximately equal within-component variances, with highly overlapping 95% CIs. This implies that individual differences shape the affective reactions to the different conditions to the same degree.

3. Bootstrapping in MLSCA

The bootstrap (Davison & Hinkley, 1997; Efron & Tibshirani, 1993) is a resampling-based method for obtaining inferential information. An estimate of the population distribution function is used to imitate the sampling process from the population. This imitation yields the bootstrap distribution of the statistic(s) of interest, from which inferential information is derived, like confidence intervals. Although the principles are straightforward, they can be implemented in various ways, yielding different bootstrap procedures. Decisions have to be made on the estimated population distribution function, the parameters of interest, the method to set up the bootstrap distribution(s), and the method to derive the required inferential information. In this section we discuss these topics from the MLSCA perspective.

3.1. The choice of the estimated population distribution function

Approaches to obtaining the estimated population distribution function (PopDF) differ in the extent to which they rely on model assumptions. No model assumptions are made in the non-parametric approach, resting on the notion that the PopDF and the empirical distribution function (EDF) converge as the sample size increases to infinity. Thus, the EDF is taken as the estimated PopDF, which implies that resampling is simply based on the observed sample data. The semi-parametric approach assumes a correctly specified model, and independently and identically distributed residuals. Resampling takes place from either the observed residuals themselves, or from a particular residual distribution, of which the parameters are estimated. The parametric approach requires the strongest assumption, namely that of a particular PopDF, of which the parameters are estimated on the basis of the sample data. Adopting stronger assumptions yields more efficient estimates. However, one should be very careful in applying the (semi-)parametric approaches, because unreasonable assumptions usually lead to severely biased estimates.

In MLSCA, one has to resort to the non-parametric approach, because the residuals are not independently distributed, making the semi-parametric approach unsuitable. The lack of distributional assumptions disqualifies the parametric approach. The non-parametric approach is standard practice in bootstrapping other types of component analysis as well, such as principal component analysis (Efron & Tibshirani, 1993; Timmerman, Kiers, & Smilde, 2007) and three-way component analysis (Kiers, 2004).

3.2. The resampling scheme

The sampling process from the population is imitated by resampling from the estimated PopDF. Usually, the resampling scheme follows the sampling scheme. In the simplest case, one has a sample from a single population. Then, the so-called bootstrap

sample is obtained by resampling n units of observation from the original sample of size n, where resampling is drawing with replacement. In MLSCA, the key question for selecting a proper resampling scheme is from which population(s) the units are considered to be sampled, or to put it another way, which level(s) are considered random and which are considered fixed. Obviously, inferential information is required only if at least one level is considered random. We discuss the three possibilities for the two-level case, and the associated resampling schemes (see Table 2 for an overview).

Table 2. Overview of resampling schemes in two-level MLSCA

Case	Fixed and random levels	Type of resampling
Multi-group	Level 2 fixed Level 1 random	Stratified resampling
Multi-observation	Level 2 random Level 1 fixed	Group resampling
Multi-level	Levels 2 and 1 random	Group resampling or double resampling

Firstly, the 'multi-group' case (Jöreskog, 1971) arises when the level 2 units are considered to be fixed, and the level 1 units random. An example of the multi-group case is when a particular patient population is to be compared with a healthy population, using random samples from both populations. A second example comes from cross-cultural research, where inhabitants are randomly drawn from a number of selected countries. The resampling scheme for the multi-group case is straightforward. The bootstrap sample is obtained using stratified resampling, where each of the group samples corresponds to a stratum. Hence, each group is resampled separately, and the resulting resampled data are combined (Davison & Hinkley, 1997, pp. 71–76).

Secondly, the level 2 units may be considered to be random, and the level 1 units to be fixed. For example, in assessing manager performance, the managers may be randomly sampled, and all employees of the selected managers are included in the study. In this case, a random level 1 interpretation is awkward because then the employees of each manager are viewed as a random sample from the population of all conceivable employees of this manager. Because this case typically occurs when an individual (e.g. manager) is observed by a number of others, we refer to this as the 'multi-observation case' in what follows. In the multi-observation case, the bootstrap sample is formed by 'group resampling', which implies resampling the level 2 units only, thereby keeping all level 1 units associated with the selected level 2 units.

Finally, the 'multi-level case' arises if the units at both levels 2 and 1 are sampled randomly from populations. For example, in examining certain patient characteristics, hospitals may be randomly sampled from the population of hospitals, and patients may be randomly sampled from each selected hospital. The non-parametric resampling scheme for the multi-level case is less straightforward than for the multi-group and multi-observation cases. Davison and Hinkley (1997, pp. 100–102) discuss two resampling strategies for two-level data, namely group resampling (see above) and 'double resampling'. Double resampling implies a two-stage resampling procedure, in which first the level 2 units are resampled, and subsequently the level 1 units are resampled within the selected level 2 units. Davison and Hinkley compare both resampling strategies for a random effects ANOVA model with balanced data.

They show that group resampling yields an expected bootstrap between group variance closer to the population between group variance than double resampling. Therefore, group resampling is preferable over double resampling for the random effects ANOVA with balanced data. For more complex situations, Davison and Hinkley suggest a (semi-)parametric approach.

MLSCA in the multi-level case certainly is more complex than the random effect ANOVA model for balanced data. MLSCA parameters are a function of the between and within covariance matrix, rather than the between and within variances only. Also, MLSCA is usually performed on unbalanced data rather than balanced data. For three MLSCA variants (i.e. all except for MLSCA-ECP), the within group covariance matrices are allowed to vary over groups. This is in contrast to the random effects ANOVA, where the within variances are assumed to be equal across groups. Thus far, multi-level bootstrapping has been used in the context of random coefficient models. Indeed, most authors used a (semi-)parametric approach (like Brumback & Rice, 1998; Carpenter, Goldstein, & Rasbash, 2003). In MLSCA, one has to resort to non-parametric resampling, but it is unclear whether the group resampling or double resampling scheme is to be preferred.

3.3. Parameters of interest

The potential parameters of interest for which inferential information is required depend on the MLSCA variant used, and whether level 2 is considered fixed or random, as is summarized in Table 3 and discussed now. In all cases, the elements of the between and within loading matrices are usually of primary interest. In the multi-group case (fixed level 2), the between component scores may be relevant, as in the empirical example in Section 2.2. Finally, the (co)variances of the within component scores may be considered.

Table 3. Parameters of interest for the MLSCA variants in the multi-group, multi-observation, and multi-level cases; $\mathbf{F}_{iw}$ denotes the within component score matrix of group i, var and cov denote variance and covariance, respectively, and (co)var($\mathbf{F}_w$) denotes the within component scores (co)variances that are constrained to be equal across I groups

Case	MLSCA-P	MLSCA-PF2	MLSCA-IND	MLSCA-ECP
Multi-group (level 2 fixed, level 1 random)	$\mathbf{B}_b$, $\mathbf{B}_w$, $\mathbf{f}_{ib}$ var($\mathbf{F}_{iw}$), $i = 1,\ldots,I$ covar($\mathbf{F}_{iw}$) $i = 1,\ldots,I$	$\mathbf{B}_b$, $\mathbf{B}_w$, $\mathbf{f}_{ib}$ var($\mathbf{F}_{iw}$), $i = 1,\ldots,I$ covar($\mathbf{F}_w$)	$\mathbf{B}_b$, $\mathbf{B}_w$, $\mathbf{f}_{ib}$ var($\mathbf{F}_{iw}$), $i = 1,\ldots,I$	$\mathbf{B}_b$, $\mathbf{B}_w$, $\mathbf{f}_{ib}$ var($\mathbf{F}_w$) covar($\mathbf{F}_w$)
Multi-observation and multi-level (both level 2 random)	$\mathbf{B}_b$, $\mathbf{B}_w$ Variance of var($\mathbf{F}_{iw}$) Variance of covar($\mathbf{F}_{iw}$)	$\mathbf{B}_b$, $\mathbf{B}_w$ Variance of var($\mathbf{F}_{iw}$) covar($\mathbf{F}_w$)	$\mathbf{B}_b$, $\mathbf{B}_w$ Variance of var($\mathbf{F}_{iw}$)	$\mathbf{B}_b$, $\mathbf{B}_w$ var($\mathbf{F}_w$) covar($\mathbf{F}_w$)

The MLSCA variants differ in their constraints on the covariance matrices of the within component scores ($\mathbf{F}_{iw}$) of the I level 2 units. As far as within (co)variances are required to be equal across level 2 units, those within (co)variances themselves are the parameters of interest, irrespective of whether the level 2 units are considered fixed or random. For example, in the most constrained model, MLSCA-ECP, the within covariances and the within variances are the parameters of interest, because both are constrained to be equal across level 2 units.

For (co)variances that are allowed to vary over level 2 units, the view taken on level 2 matters. If level 2 is considered fixed, the (co)variances for each observed level 2 unit are to be interpreted. In the random case, the distribution(s) of the (co)variance(s) are to be characterized in some way. To this end, we use the variance of the (co)variances, which characterizes the spread of the (co)variances across level 2 units. Note that the mean of the distribution is not of interest, as it is fixed for identification purposes (at 0 for covariances and at 1 for variances).

3.4. Setting up the bootstrap distributions

Once the parameters of interest have been determined, a method for setting up the bootstrap distributions should be chosen. In so doing , it is very important to account for possible non-uniqueness in the parameter estimates, because otherwise the bootstrap distributions are arbitrarily inflated (Milan & Whittaker, 1995). The loading matrices in all MLSCA variants are non-unique. Their degree of transformational freedom differs, and we start by discussing the ones with largest degree of transformational freedom.

Transformational freedom is found in the between loading matrix of all MLSCA variants, and in the within loading matrices of the MLSCA-ECP and MLSCA-P models. Because the transformations of the between and within loading matrices are independent, they are performed separately. The transformational freedom in MLSCA loadings can be treated analogously as in principal component analysis (PCA). To what extent the user exploits the transformational freedom should be reflected in the CIs for PCA loadings, as argued by Timmerman, Kiers, and Smilde (2007). For example, if one considers the principal axes solution, 95% CIs for the principal axes loadings should cover the population principal axes loadings in 95% of the estimated CIs. This is achieved by selecting a proper method to set up the bootstrap distribution. In analogy to Timmerman, Kiers, and Smilde (2007), we will discuss the 'fixed criterion', 'best interpretable', and 'target' cases.

The 'fixed criterion' case implies that one uses a fixed rotation criterion, determined in advance, such as varimax. To obtain bootstrap CIs that cover the varimax-rotated population loadings with the required probability, each bootstrap loading matrix should be varimax rotated as well. As such, rotated solutions are unique up to reflection and permutation only, and the rotated bootstrap loading matrices should be made comparable by reflecting and reordering each rotated bootstrap loading matrix so that they match each other optimally (Lambert, Wildt, & Durand, 1991; Milan & Whittaker, 1995). This can be done by maximizing the sum of congruence coefficients (Tucker, 1951) between the sample loading matrix and each bootstrap loading matrix, over all possible permutations of the columns of the bootstrap loading matrix. Subsequently, the bootstrap distributions of the elements of the loading matrix can be readily obtained.

As an alternative to the fixed criterion case, one may inspect various rotated solutions and select the 'best interpretable' sample loading matrix. For that case, Timmerman, Kiers, and Smilde (2007) recommend Procrustes rotation of the bootstrap loadings towards the selected loading matrix of the original sample. One could use oblique Procrustes rotation, or orthogonal Procrustes rotation if one excludes oblique rotation in advance.

Procrustes rotation of bootstrap loadings is also used in the 'target case', where a particular target loading matrix is available. Because Procrustes rotated loadings are unique, the bootstrap distributions can be readily obtained.

The quality of bootstrap CIs in relation to the rotation performed has been examined in PCA (Timmerman, Kiers, & Smilde, 2007). Because MLSCA is intimately related to

PCA, it is to be expected that those results apply to MLSCA as well. The CI quality of the best interpretable and target cases, which involve Procrustes rotations in the bootstrap procedure, appeared to be unaffected by the rotation performed. However, the fixed criterion case appears to suffer from low bootstrap CI quality in a particular condition, namely with highly unstable rotated loadings over samples. The latter occurs when the position of the axes in terms of the fixed criterion is arbitrary, for example when the variables have a circumflex structure (meaning that they can be ordered along a semicircle), and one of the usual simplicity criteria, such as varimax, is applied (Timmerman, Kiers, & Smilde, 2007).

The within loading matrices of MLSCA-PF2 and MLSCA-IND are unique up to permutation and reflection. Therefore, the bootstrap loading matrices resulting from those methods should be reflected and reordered so that they match each other optimally before the bootstrap distributions are obtained. This can be done by, for example, maximizing the sum of congruence coefficients between columns of the bootstrap loading matrices and the sample loading matrix.

Note that once the non-uniqueness of the between and within loading matrices has been resolved, the associated component scores are unique as well. Thus, bootstrap distributions of parameters related to those component scores can be obtained directly.

3.5. Confidence intervals from the bootstrap distributions

Various methods exist to estimate a CI from the bootstrap distribution (Davison & Hinkley, 1997; Efron & Tibshirani, 1993). They are based either on the bootstrap standard error (the standard deviation of the bootstrap distribution) or on the percentiles of the bootstrap distribution. The percentile-based methods have the advantage of being transformation respecting and range preserving. Moreover, the quality in terms of coverage of the percentile-based bootstrap CIs in PCA appeared consistently higher than that of a bootstrap standard error-based method in a comparative simulation study (Timmerman, Kiers, & Smilde, 2007). Therefore, we will focus on the percentile-based methods.

The simplest method is the so-called percentile method, which uses the central $(1 - 2\alpha)$ part of the cumulative bootstrap distribution as the approximate central $100(1 - 2\alpha)\%$ confidence interval. Thus, the lower and upper endpoints are given by $\hat{\theta}^*(\alpha)$ and $\hat{\theta}^*(1 - \alpha)$, respectively, where $\hat{\theta}^*(\alpha)$ denotes the 100αth percentile of the bootstrap distribution. The percentile method works well in the absence of bias (i.e. the centres of the sampling and bootstrap distributions coincide) and in case of constant variance of the sampling distribution. Otherwise, the BCa method (Efron, 1987) usually achieves better performance using adapted percentiles of the bootstrap distribution. The adapted percentiles depend on a bias and an acceleration correction. The estimated bias correction, denoted by $\hat{z}_0$, corrects for the discrepancy between the centres of the bootstrap and sampling distributions. The estimated acceleration, denoted by $\hat{a}$, corrects for violation of the assumption that the standard error of the statistic $\hat{\theta}$ is constant over all values of the parameter θ. The BCa method estimates the lower and upper endpoints of the $100(1 - 2\alpha)\%$ CI by $\hat{\theta}^*(\alpha_1)$ and $\hat{\theta}^*(\alpha_2)$, where

$$\alpha_1 = \mathbf{\Phi}\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{\alpha}}{1 - \hat{a}(\hat{z}_0 + z^{\alpha})}\right), \qquad \alpha_2 = \mathbf{\Phi}\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right), \tag{2}$$

with $\mathbf{\Phi}$ the standard normal cumulative distribution function, and z^{α} the 100αth percentile point of the standard normal distribution.

The bias correction can be estimated from the proportion of the bootstrap estimators that is less than the sample estimate $\hat{\theta}$, namely as

$$\hat{z}_0 = \mathbf{\Phi}^{-1}\left(\frac{\#(\hat{\theta}_b^* < \hat{\theta})}{B}\right), \tag{3}$$

where $\mathbf{\Phi}^{-1}$ is the inverse of $\mathbf{\Phi}$.

The computation of the acceleration is based on the so-called empirical influence values. Empirical influence values, to be estimated for each unit of observation, can be obtained in various ways (Davison & Hinkley, 1997). We use the easily computed jackknife approximation, which we will first discuss for the general multi-sample case. The jackknife empirical influence value l_{k_i} (Davison & Hinkley, 1997, p. 76) for individual k_i in group i is given by

$$l_{k_i} = (K_i - 1)(\hat{\theta} - \hat{\theta}_{-k_i}), \tag{4}$$

where $\hat{\theta}$ denotes the sample parameter estimate and $\hat{\theta}_{-k_i}$ denotes the parameter estimate obtained by omitting case k_i from the sample. The multi-sample acceleration estimate (Davison & Hinkley, 1997, p. 210) is given by

$$\hat{a} = \frac{1}{6}\left(\frac{\sum_{i=1}^{I} K_i^{-3}\sum_{k_i=1}^{K_i} l_{k_i}^3}{\left(\sum_{i=1}^{I} K_i^{-2}\sum_{k_i=1}^{K_i} l_{k_i}^2\right)^{3/2}}\right). \tag{5}$$

The single-sample acceleration estimate is a special case of the multi-sample estimate, with $I = 1$.

The specific acceleration estimation procedure in MLSCA depends on the resampling scheme used. In the case of stratified resampling and double resampling, we use the multi-sample acceleration estimate in (5). In the case of group resampling, we use the single-sample acceleration estimate version of (5), but now considering the level 2 units ($i = 1, \ldots, I$) as the units of observation, rather than the level 1 units ($k_i = 1, \ldots, K_i$). Thus, the acceleration is estimated on I empirical influence values, which are computed by successively omitting each of the I groups from the data.

4. Simulation study

In the previous section, we discussed procedures for estimating bootstrap CIs for parameters in the MLSCA variants, where we distinguished the multi-group, multi-observation, and multi-level cases. To gain an insight into the quality of the resulting CIs, we performed a simulation study, covering the multi-group, multi-observation, and multi-level cases. In particular, we questioned to what extent the BCa CIs outperform the percentile ones, which sample size(s) are needed to obtain reasonable quality, and, for the multi-level case, whether group resampling or double resampling is to be preferred.

To keep the study feasible we made a number of limiting choices. Firstly, we confined ourselves to the least constrained MLSCA variant, MLSCA-P, which takes by far the least computational effort of the four variants. Thus, some insight is obtained into the CI quality of between and within loadings and the (co)variances of within

component scores. Secondly, to generate the data, used only population loading matrices with a simple structure and applied a fixed rotation criterion (normalized varimax) to the sample between and within loading matrices. The effects of this choice are known from results in PCA (Timmerman, Kiers, & Smilde, 2007): with simple structure population loadings, the fixed rotation and Procrustes rotation yield the same similar CI qualities; with less simple structures the CI quality using Procrustes rotations remains good, but deteriorates using fixed rotation with one of the usual simplicity criteria.

4.1. The simulated data and their analyses

A natural approach to simulating MLSCA-P data is to sample from population data constructed according to equation (1). By choosing the population offsets ($\mathbf{m}$) to be 0, a simulated sample data matrix $\mathbf{Ysim}_i$ for the ith group was constructed as

$$\mathbf{Ysim}_i = \mathbf{1}_{K_i}\mathbf{f}'_{ib}\mathbf{B}'_b + \mathbf{F}_{iw}\mathbf{B}'_w + \mathbf{E}_i. \tag{6}$$

where $\mathbf{f}_{ib}$ is the vector of between component scores of group i, $\mathbf{F}_{iw}$ the within component scores matrix of group i, $\mathbf{B}_b$ the between loading matrix, $\mathbf{B}_w$ the within loading matrix, and $\mathbf{E}_i$ the error matrix.

The expected proportions of variance as contributed by the between, within, and error parts of the simulated data were kept fixed at .15, .60, and .25, respectively. In a preliminary study, varying this distribution (within reasonable bounds, i.e. no proportion smaller than .15) hardly influenced the results.

We used a fixed number of nine variables, three between components and two within components. The between and within loading matrices were constructed to be optimally simple in terms of the varimax criterion, namely as

$$\mathbf{B}_b = \begin{bmatrix} c_b & 0 & 0 \\ c_b & 0 & 0 \\ c_b & 0 & 0 \\ 0 & c_b & 0 \\ 0 & c_b & 0 \\ 0 & c_b & 0 \\ 0 & 0 & c_b \\ 0 & 0 & c_b \\ 0 & 0 & c_b \end{bmatrix} \quad \text{and} \quad \mathbf{B}_w \begin{bmatrix} c_w & 0 \\ c_w & 0 \\ c_w & 0 \\ c_w & 0 \\ c_w & 0 \\ 0 & c_w \\ 0 & c_w \\ 0 & c_w \\ 0 & c_w \end{bmatrix},$$

where c_b and c_w are constants, chosen such that the required expected proportion of structural variance for the between and within parts is obtained: $c_b = .38$ and $c_w = .77$.

The shape of the distribution of the between and within component scores and the elements of the error matrix was varied at two levels. We used a multivariate normal distribution and a symmetric, leptokurtic distribution (E(skewness) = 0, E(kurtosis) = 5), using the procedure proposed by Ramberg, Dudewicz, Tadikamalla, and Mykytka (1979), both with covariance matrix as specified below. Those two distributions were selected, as they showed the largest differences in quality of

bootstrap CIs in PCA (Timmerman, Kiers, & Smilde, 2007) among various distributions, and we expected those effects to be similar in MLSCA.

The simulated data for the multi-group, multi-observation, and multi-level cases differ in the covariance matrices of the component scores, $S(\mathbf{F}_b)$ and $S(\mathbf{F}_{iw})$, $i = 1, \ldots, I$. The differences boil down to a fixed or random interpretation, which is expressed in either an observed or expected covariance matrix with a particular structure.

In the MLSCA-P model, the within covariance matrices, $S(\mathbf{F}_{iw})$, are allowed to vary over the I groups, with $\sum_{i=1}^{I} S(\mathbf{F}_{iw}) = \mathbf{I}_{Q_w}$ for identification purposes, without loss of generality. In the case of a random level 2 (i.e. the multi-observation and multi-level cases), the expected within covariance matrix over the I groups, $E(S(\mathbf{F}_{iw}))$, was identity. In the multi-group case, with fixed level 2, the sum of the I simulated within covariance matrices equalled identity.

In the simulation study, for each group i, we sampled the elements of the (2×2) covariance matrix $S(\mathbf{F}_{iw})$ from uniform distributions. The diagonal elements, the variances, were sampled from the interval [0.5,1.5], and the off-diagonal elements, the covariances, from $[-0.5, 0.5]$. For the case with fixed level 2, the sampled elements were centred such that $\sum_{i=1}^{I} S(\mathbf{F}_{iw}) = \mathbf{I}$ to make them in agreement with the fixed level 2 model. Then, in the case of a random level 1 (multi-level and multi-group cases), the elements of $\mathbf{F}_{iw}$ for group i were drawn such that they had an expected covariance matrix $S(\mathbf{F}_{iw})$. For a fixed level 1 (multi-observation case), the within component scores were drawn such that they had an observed covariance matrix $S(\mathbf{F}_{iw})$.

In the case of a random level 2 (multi-observation and multi-level cases), the between component scores were drawn such that the expected between covariance matrix, $E(S(\mathbf{F}_b))$, equals identity. If level 2 is considered fixed (the multi-group case), the between component scores were drawn such that the observed between covariance matrix, $S(\mathbf{F}_b)$, equals identity.

In the experiment, we varied the number of samples. We used a pre-specified mean number of individuals per group, using a varying number of individuals within the groups, ranging from three less than to three more than the mean group size. For the multi-group case, we used a fixed number of 10 groups and mean group sizes of 20, 100, and 200. For the multi-observation case, we varied the number of groups (20, 100, and 200) only, using a mean group size of 10. For the multi-level case, we varied both the number of groups (20 and 40), and the mean group sizes (30, 100, and 200). The design with factors sample size and distribution (level levels) was completely crossed, with 1,000 replicates per cell, leading to 6,000 (multi-group case) + 6,000 (multi-observation case) + 12,000 (multi-level case) = 24,000 generated data matrices.

Each simulated data matrix was analysed by MLSCA-P, with the correct numbers of between and within components. The parameters of interest were the elements of the between and within loading matrices. Furthermore, in the multi-group case, the (co)variances of the within component scores themselves were considered, and in the multi-observation and multi-level case the variances of the variances and the covariance of the within component scores. For each parameter of interest, bootstrap 95% CIs were estimated, using both the percentile and BCa methods. The number of bootstrap samples was chosen to be 1,000. All between and within loading matrices were rotated using the normalized varimax criterion. To resolve arbitrary reflection and reordering, each rotated bootstrap loading matrix was optimally permuted and reflected to the sample solution by maximizing the sum of congruence coefficients (Tucker, 1951) between corresponding loadings of the sample and the bootstrap sample.

The quality of the estimated CIs was assessed by considering the coverage, which is computed as the percentage of the 1,000 95% CIs per condition that include the population parameter. Consequently, a coverage of 95% is the ideal value. The optimal MLSCA-P estimates of the population data are the population parameters, after normalized varimax rotation of the between and within loadings, and optimal permutation and reflection to the sample solution. The latter is necessary to account for arbitrary differences between the sample solution, on which the CIs are based and the population parameters. The population loadings were derived analytically, whereas the population variances of (co)variances and the population (co)variances were estimated numerically.

The coverage is computed for each parameter separately. This measure suffers from sampling error. The standard error of the 95% coverage for each single parameter is 0.69%, assuming a binomial distribution. For ease of presentation, the results of equivalent parameters are taken together (namely the high loadings (population loadings non-zero), low loadings (population loadings zero) and the variances of within component scores). The standard error of the combined scores is lower than the ones presented here, but cannot be computed exactly because of dependencies between parameters.

4.2. Results

The mean coverage indicates the overall quality of the 95% CIs. The mean coverages for the multi-group, the multi-observation, and the multi-level cases have been plotted in Figures 3–5, respectively, for each parameter of interest, and as a function of sample size and distribution.

In the multi-group case (see Figure 3), the coverage generally approaches the desired value of 95% with increasing group sizes, for the loadings and the (co)variance of within component scores. The coverage of CIs for data generated with component scores and errors drawn from a normal distribution is consistently slightly better than that when drawn from a leptokurtic distribution, and this difference between distributions diminishes with increasing sample size. The BCa method outperforms the percentile method for high between and within loadings, where the difference between the methods diminishes with increasing sample size. Both methods perform comparably for

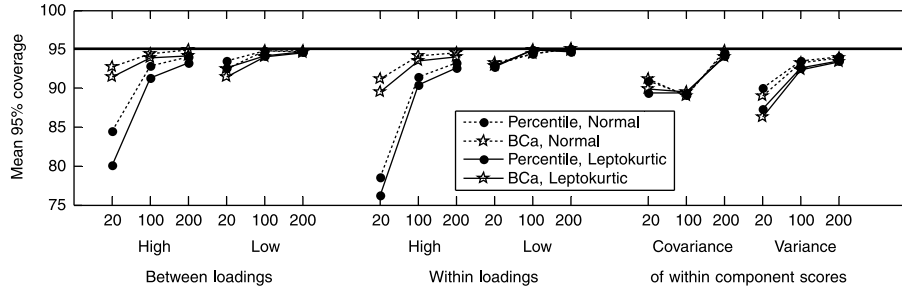

Figure 3. Multi-group case (level 2 fixed, level 1 random): mean coverage of the percentile and BCa methods, as a function of mean group size ($k = 20$, 100, and 200) and distribution of component scores and error (normal and leptokurtic), for high and low between and within loadings, and covariances and variances of within component scores.

the other parameters. The clearly superior performance of the BCa method in the case of high loadings with small group sizes is to be expected, as the bootstrap distribution in those cases is skewed, and the BCa method compensates for this skewness. The current results indicate reasonable coverage of the CIs for all parameters with minimal group sizes of about 100. For the loadings, the BCa CIs already have a reasonable quality for group sizes of 20.

In the multi-observation case, with level 2 random and level 1 fixed (see Figure 4), the mean coverage approaches 95% with increasing number of groups, for all parameters, as expected. Also, the coverage of CIs for data generated with normally distributed component scores and errors is consistently somewhat better than for leptokurtic distributed ones. The difference in coverage between distributions diminishes with increasing sample size. The performances of the BCa and percentile methods are surprisingly similar. Only in conditions with skewed bootstrap distributions (high between and within loadings) does the BCa perform slightly better. The percentile method even outperforms the BCa method somewhat in the low between loading condition, with a small number of groups ($I = 20$). This is because low loadings have symmetrical sampling distributions, and the BCa method erroneously corrects in cases with such a small sample size. Generally, a sample consisting of at least 100 groups appears necessary to obtain CIs of reasonable quality for all parameters of interest.

For the multi-level case, the mean coverage of the CIs in the condition with normally distributed component scores and errors is plotted in Figure 5, for each parameter of interest and as a function of sample size. As can be seen from Figure 5, the coverage of the CIs approaches 95% with increasing number of groups and group sizes. Overall, good quality CIs are obtained only when the sample sizes at both levels are reasonable, such as at least 40 groups with group sizes of about 100.

In the top part of Figure 5, an irregular effect of sample size is seen for the loadings, both at the between and within level. The effect is rather salient for the high loadings in the condition with 20 groups, and becomes less pronounced with 40 groups. This counter-intuitive finding is due to the dependency of the (estimated) EDFs of the between and the within part. Increasing the group size leads to a better estimate of the within part of the groups in the sample. However, when the number of groups is small, the chances are rather high that the groups in the sample are not representative of the PopDFs of the population of groups. Thus, the effects for the parameters of the between part are much more pronounced than for those of the within part. This is because the

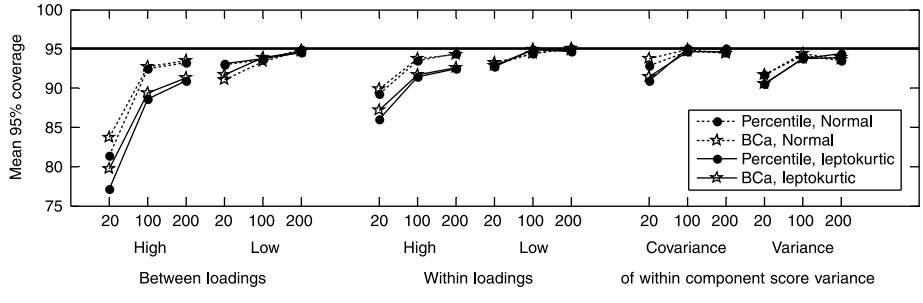

Figure 4. Multi-observation case (level 2 random, level 1 fixed): mean coverage of the Percentile and BCa methods, as a function of number of groups ($I = 20$, 100, and 200), and distribution of component scores and error (normal and leptokurtic), for the high and low between and within loadings, and the covariance and variance of within component scores.

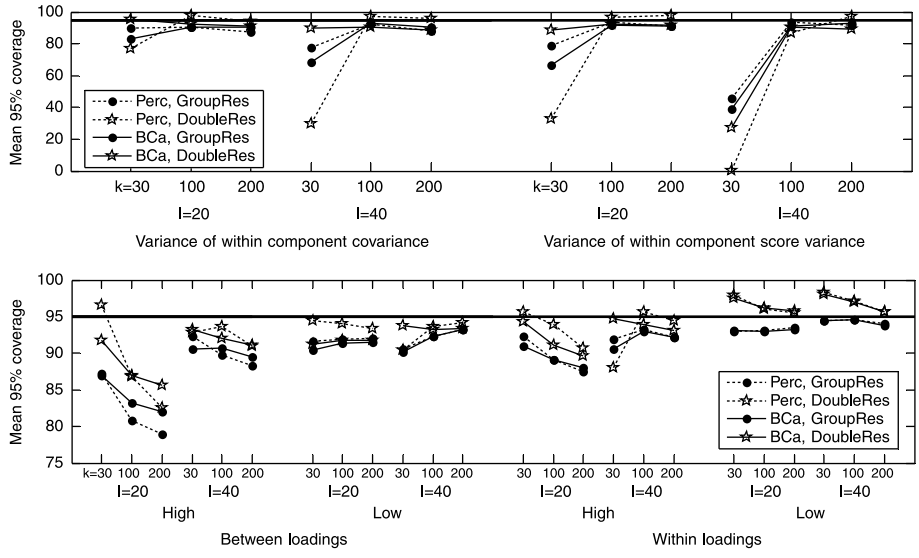

Figure 5. Multi-level case: mean coverage of the percentile and BCa methods, as a function of number of groups ($I = 20, 40$) and mean group size ($k = 30$, 100, and 200) for normally distributed component scores and errors, for the high and low between and within loadings (top), and the covariance and variance of within component scores (bottom).

within parts vary relatively little over groups: the basis for each within part (within loading matrix) equals across groups, and only the within component score (co)variances differ.

The resampling strategy and the CI type clearly influence the coverage. For the loadings, the double resampling approach almost consistently results in CIs with a higher coverage than group resampling. Overall, the double resampling BCa intervals appear closest to the desired 95% coverage and are therefore preferred.

Thus far, we discussed the results of the simulation study for the multi-level case as summarized in Figure 5, that is, the condition with normally distributed component scores and errors. The pattern of differences between the normal and leptokurtic conditions appeared similar to those found in the multi-group and multi-observation cases, and therefore those results are not plotted here. The similar pattern in differences implies that the coverage of the CIs in the normal distribution condition is consistently somewhat better than in the leptokurtic condition, and this difference diminishes with increasing sample sizes; the difference in coverage between the distributions becomes negligible in conditions with reasonable sample sizes at both levels (i.e. minimally 40 groups with group sizes of about 100).

5. Conclusion

This paper has presented bootstrap strategies for estimating confidence intervals in MLSCA. Herewith, the main choices to be made involve the selection of the resampling scheme and the procedure to set up the bootstrap distributions.

The proper resampling scheme depends on which levels are considered to be random and fixed. For multi-level data consisting of two levels, we distinguished the

multi-group (only level 1 random), multi-observation (only level 2 random), and multi-level cases (both levels random). It was argued that, on a theoretical basis, the proper resampling scheme for the multi-group case appears to be stratified resampling and for the multi-observation case group resampling. For the multi-level case, group resampling or double resampling may be appropriate.

In a simulation study, the quality of bootstrap CIs for parameters of the MLSCA-P variant was assessed for the three cases. If only one level is considered random, that is, in the multi-group and multi-observation cases, good quality CIs appeared to be obtained for all parameters with sample sizes of about 100 (i.e. 100 individuals per group, and 100 groups, respectively). Smaller sample sizes result in CIs that are clearly too small.

For the multi-level case, two types of potentially good resampling strategies were compared, namely group resampling and double resampling. Double resampling generally appeared to yield better CIs than group resampling, and hence appears to be the preferred strategy for MLSCA. The preference for double resampling rather than group resampling contrasts with the preferred resampling strategy for a random effects ANOVA with balanced data (Davison & Hinkley, 1997; pp. 100–102). Furthermore, CIs with reasonable coverage are obtained when the sample sizes at both levels are sufficiently large. This is not surprising, as one generalizes to both levels of populations. The results of our simulation study suggest that, say, 40 groups with group sizes of about 100 are a minimal requirement to achieve proper CIs.

Overall, the BCa confidence intervals appeared to be closer to the desired coverage than the percentile method, and hence the BCa method is the most recommendable. This finding is not very surprising, as the BCa method is second-order correct, and the percentile method only first-order correct (Efron & Tibshirani, 1993).

The selection of a procedure to set up the bootstrap distributions should account for the possible non-uniqueness in the parameter estimates. As discussed, the degree of non-uniqueness depends on the MLSCA variant at hand. The procedure to set up the bootstrap distributions is rather straightforward for MLSCA variants that are unique up to permutation and reflection. However, for variants with transformational freedom (MLSCA, MLSCA-P, and MLSCA-ECP) the procedure selected should reflect the extent to which the user exploits the transformational freedom. We have discussed three cases, namely the fixed criterion, best interpretable and target cases, and their associated proper methods to set up the bootstrap distribution. The latter two cases involve Procrustes rotations in the bootstrap procedure, and the CI quality is not affected by the rotation performed. However, the fixed criterion case may suffer from bad CI quality, namely when rotated loadings are highly unstable over samples.

The availability of CIs for MLSCA parameters[1] aids a proper interpretation of the analysis results. This is illustrated by our empirical example from an experimental social dilemma study. In this study, the participants were randomly allocated to one of the six experimental conditions. Because the conditions are considered to be fixed, this yields a multi-group case, and stratified resampling was used in the bootstrap procedure. The orientation of the axes of the sample loadings was chosen such that the orthogonally rotated solution was best interpretable, and therefore the bootstrap loadings were orthogonally Procrustes rotated towards the sample loadings. The CIs of the between and within loadings revealed which emotions are significantly attached to

[1] *Matlab code for estimating the various CIs discussed here can be obtained from the first author.*

which components. Moreover, the CIs associated with the between component scores identify evidence for the presence of main and interaction effects of the experimental conditions.

The different strategies for obtaining CIs stress the necessity to explicitly define the parameters and the population(s) of interest. The latter implies that it is to be decided for each level whether it is considered fixed or random. When more than a single level is considered to be random, the number of observations necessary to obtain reliable inferential information increases dramatically. Therefore, in designing a study, it is important to assess for each level whether a random interpretation is strictly necessary. In practice, often only a limited number of observations can be made. Then, a study using a compromise of limited sample sizes for all levels considered random may easily reveal a disappointing amount of information about all populations under study. Instead, it may be more informative to focus on a single random level, and to use the available observations to achieve a reasonable sample size at this very level.

Acknowledgements

The authors thank two anonymous reviewers, and Thom Baguley and Johan A. Westerhuis for their comments on an earlier version of this paper.

References

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230–258.

Brumback, B. A., & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, *93*, 961–976.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models, applications and data analysis methods*. Newbury Park, CA: Sage.

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics*, *52*, 431–443.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

De Noord, O. E., & Theobald, E. H. (2005). Multilevel component analysis and multilevel PLS of chemical process data. *Journal of Chemometrics*, *19*, 301–307.

Du Toit, S. H. C., & Du Toit, M. (2008). Multilevel structural equation modeling. In J. De Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis*. New York: Springer.

Efron, E. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185.

Efron, E., & Tibshirani, R. J. (1993). *An introduction into the bootstrap*. New York: Chapman & Hall.

Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., & Smilde, A. K. (2005). Multilevel component analysis of time-resolved metabolic fingerprinting. *Analytica Chimica Acta*, *530*, 173–183.

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

Kiers, H. A. L. (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics*, *18*, 22–36.

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

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

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, *26*, 421–434.

Milan, L., & Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, *44*, 31–49.

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

Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., & Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics*, *21*, 201–214.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, *59*, 301–320.

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.

Timmerman, M. E., Kiers, H. A. L., & Smilde, A. K. (2007). Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, *60*, 295–314.

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

Received 12 March 2007; revised version received 20 November 2007