Check for updates

Received: 7 October 2019 | Revised: 19 February 2020 | Accepted: 3 March 2020

DOI: 10.1002/cem.3233

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Detecting outlying variables in multigroup data: A comparison of different loading similarity coefficients

Sopiko Gvaladze[1] | **Kim De Roover**[1,2] | **Francis Tuerlinckx**[1] | **Eva Ceulemans**[1]

[1]Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

[2]Department of Methodology, Tilburg University, Tilburg, The Netherlands

Correspondence
Sopiko Gvaladze, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Flanders, Belgium.
Email: sofia.gvaladze@kuleuven.be

Funding information
Flemish Government—department EWI; Hercules Foundation; Research Fund of KU Leuven, Grant/Award Number: C14/19/054

Abstract

Multivariate multigroup data are collected in many fields of science, where the so-called groups pertain to, for instance, experimental groups or countries the participants are nested in. To summarize the main information in such data, principal component analysis (PCA) is highly popular. PCA reduces the variables to a few components that are linear combinations of the original variables. Researchers usually assume those components to be the same across the groups and aim to apply a simultaneous component analysis. To investigate whether this assumption is reasonable, one often analyzes the groups separately and computes a similarity index between the group-specific component loadings of the variables. In many cases, however, most variables have highly similar loadings across the groups, but a few variables, which we will call "outlying variables," behave differently, indicating that a simultaneous analysis is not warranted. In such cases, the outlying variables should be removed before proceeding with the simultaneous analysis. To do so, the variables are ranked according to their relative outlyingness. Although some procedures have been proposed that yield such an outlyingness ranking, they might not be optimal, because they all rely on the same choice of similarity coefficient without evaluating other alternatives. In this paper, we give an overview of other options and report extensive simulations in which we investigate how this choice affects the correctness of the outlyingness ranking. We also illustrate the added value of the outlying variable approach by means of sensometric data on different bread samples.

KEYWORDS

component similarity, multivariate multigroup data, PCA, SCA, Tucker's congruence

1 | INTRODUCTION

In many research areas, including disciplines like chemistry and psychology, studies are conducted that yield linked or coupled data blocks.[1,2] The main feature of this kind of data is that the blocks share one of the two modes, that is, either the column or variable mode (i.e., column-coupled data or multigroup data) or the row or object mode (i.e., row-coupled data or multiblock data). In this paper, we will focus on column-coupled data. For a first example, one can think of studies that inspect the expression of multiple genes in a few participant groups (e.g., groups that are vaccinated with different vaccines[3]). A second example is given by cross-cultural studies in which inhabitants of different

 © 2020 John Wiley & Sons, Ltd.

countries fill in the same questionnaire.[4] A third example that we will focus on in this paper is sensory profiling data, where a number of assessors rate different food samples with respect to a set of sensory attributes.[5,6]

When analyzing column-coupled multigroup data, extracting constructs that summarize the linear relations between the variables in all groups simultaneously is often of prime interest.[7,8] To this end, principal component analysis (PCA) and multigroup variants thereof, such as simultaneous component analysis (SCA),[8] are highly popular across different disciplines. Applying SCA to the vertically concatenated data of the different groups yields a few components and a decomposition of the original data into component scores and loading matrices. The loading matrix, which is identical for each group, reflects the linear relationships between the variables and the components and thus allows to interpret and label the constructs captured by the components. The score matrix holds the scores of the objects on the components and differs across the groups.

In psychometrics, after the SCA analysis is conducted, researchers often execute follow-up analyses in which they investigate whether the groups differ with respect to their mean scores on the obtained SCA constructs.[4] Alternatively, they inspect how the obtained SCA components relate to other measured variables by computing correlations between the component scores and these covariates.[9] Such follow-up analyses imply the crucial assumption that the variables behave similarly across the groups and thus that it makes sense to force the loadings and thus the components to be identical across the groups.[10–13] If this assumption does not hold, the constructs captured by the SCA components cannot be interpreted in a meaningful way across the groups as they will strike a middle ground between the group-specific solutions, and the follow-up analyses regarding these constructs are shaky at best. Thus, though parsimonious and simple, this SCA-based approach may not be satisfactory if some of the variables behave differently across the groups in that they have different linear relationships with the components. This phenomenon is well known in the psychometric literature, where it is called differential item functioning.[14,15] At this point, the researcher has two options: the first option is to switch from using SCA to running separate PCA analyses per group, and the second option is to check for and remove variables that have different correlation patterns across the groups. If most of the variables have similar correlation patterns, the latter option is often more desirable, because the SCA model is more parsimonious than separate PCA models and because we gain more insight into which variables show systematic correlational differences between the groups.[16] We will call these variables "outlying variables." Note that these variables should not be confused with outlying objects (i.e., multivariate outliers, which are objects that have deviating scores on the variables[17]) or with variables that hardly correlate with any of the other variables.[18]

Although less extensively discussed, similar concerns exist in chemometrics and sensory profiling. Næs et al elaborate on the importance of evaluating whether different assessors use sensory variables in a similar way.[19] Because these assessors are extensively trained and calibrated beforehand, individual differences in the correlation structure of the variables are considered nuisance effects, and the variables causing it should be removed. To do so, they build, for instance, on the work of Dahl et al[20] and propose to fit SCA models (i.e., the assessors constitute the groups) with shared scores rather than loadings. Next, they plot the component loadings of one single variable at a time using labels to indicate the different assessors. This way, it can be visually assessed whether the loadings of an attribute differ considerably across the assessors, indicating that it may be wise to remove this attribute. Moreover, De Roover et al[21] and Wilderjans et al[22] mention the strict assumptions underlying the Candecomp/PARAFAC (CP) model[23–25] that is very popular in chemometrics. CP is a component analysis approach for three-mode data. These data are a specific case of column-coupled data, in which the groups share not only the column mode but also the row mode. A CP analysis extracts a set of components from the data, on which the elements of all three modes score. Thus, CP imposes the assumption that the component loadings are the same in all the groups. This assumption may not always be reasonable. Again, it can be the case that the structures within the different groups are so widely different that applying a different technique (e.g., clusterwise PARAFAC, SCA, or separate PCAs) is recommended. However, in other cases, the differences may be concentrated around one or a few variables that behave differently across the group. In such cases, it can be recommended to detect and remove variables that violate the CP assumption, so that CP can still be used. Moreover, inspecting the removed variables can yield useful information about the data at hand.

Within psychometrics, De Roover et al[26] and Gvaladze et al[27] recently developed and evaluated multiple heuristics to do just that. All these heuristics imply that one runs separate PCA analyses on the different groups, performs some kind of oblique Procrustes transformation (also known as rotation), and finally assesses the similarity of the resulting loadings by means of Tucker's congruence index.[28–31] On the basis of the similarity results, the heuristics yield an outlyingness ranking of the variables. To obtain this ranking, first, one computes for each variable how much the congruence value would increase upon its removal. The variable for which this increase is the largest is the most outlying variable. Next, this most outlying variable is removed, the PCA solutions are recomputed and transformed

(i.e., rotated), and the second most outlying variable is identified and removed in the same way and so on until no more variables can be removed.* Given this outlyingness ranking, the heuristics implement a different stopping rule to decide how many variables are outlying. Overall, on the basis of the simulation results reported by De Roover et al[26] and Gvaladze et al,[27] the methods look promising, however.

A first aim of this paper is to introduce this research line on outlying variable detection in column-coupled data to the field of chemometrics and sensory profiling. As we have discussed above, establishing whether attributes are used in the same way by assessors is an important topic in sensory profiling. Furthermore, the assumptions of the very popular CP model are quite strict, and by loosening and questioning these assumptions, one might gain further insight into the data at hand.

A second aim of this paper is to further investigate the outlyingness ranking procedure. The results of De Roover et al[26] and Gvaladze et al[27] are incomplete because they only evaluated the Tucker's congruence index, whereas many other indices have been proposed in the literature: the *Rv* coefficient[28,32,33] and its modified variant,[34] the angle between the subspaces formed by the components,[35,36] and the root-mean-square difference (RMSD) between the loadings.[37,38] On the one hand, these measures differ in which loadings are involved in the computation of the index (all loadings, loadings on a specific construct, and loadings of a specific variable). These alternative loading selections correspond nicely to different ways of looking at loading differences and similarities (e.g., applying a similarity measure per variable is related to inspecting for each variable whether it has the highest loading on the same component in all groups). On the other hand, some of the coefficients are affected by all component transformations, whereas others are not, or only by oblique transformations. If transformation-invariant indices would perform equally well as indices that are sensitive to transformations, this would be important to know, because many transformation criteria exist[39–43] adding to the arbitrariness of the analysis. Thus, the second aim of this paper is to shed light on whether and how the choice made among the similarity indices affects the outlyingness ranking. To this end, we run an extensive simulation where we simulate data with a few outlying variables and investigate whether these variables constitute the top of the outlyingness ranking. If this is not the case, no ranking-based heuristic can ever yield the correct set of outlying variables. We will disregard the performance of subsequent stopping criteria to determine the number of outlying variables, because multiple stopping criteria have been developed and considering them all would make the endeavor considerably more complex. Moreover, these criteria build on the Tucker's congruence score and thus may have to be adjusted before applying them to other similarity indices. We think that such adjusting is only worthwhile, if other similarity indices would perform better than Tucker's congruence in terms of correctly ranking the variables. Note furthermore that we focus on data that consist of two groups. If the data consist of more than two groups, the procedure can be applied in a pairwise manner or on clusters of groups that are induced from the data, using a clustering approach that focuses on loading differences.[44,45]

The remainder of this paper is organized as follows. First, we recapitulate the data structure and preprocessing in Section 2. In Section 3, we recapitulate the PCA-based approaches for analyzing column-coupled data. Section 4 discusses how the outlyingness ranking is obtained in the above-mentioned heuristics,[26,27] followed by a brief discussion of the associated limitations. Section 5 recapitulates the different similarity measures that can be plugged into the heuristics as an alternative to Tucker's congruence. Section 6 presents an extensive simulation study investigating which similarity measures provide the user with a correct outlyingness ranking. In Section 7, we apply the outlyingness ranking heuristic to an empirical data set on sensory bread ratings. In Section 8, we discuss our findings and directions for future work.

2 | DATA STRUCTURE AND PREPROCESSING

In this paper, we assume that the data stem from two groups and thus comprise two data blocks, where each data block $\mathbf{X}_i$ ($i = 1,2$) consists of the scores of N_i objects on the same J variables. The number of objects in each group N_i is allowed to differ between groups. The full data set $\mathbf{X}$ is obtained by vertically concatenating the $\mathbf{X}_i$ blocks and thus counts $N = N_1+N_2$ rows. Moreover, we assume that J is considerably smaller than N. In case J approaches or exceeds N, the heuristics discussed in this paper probably would not perform well, because the component solutions would become unstable, and in this case, other options like sparse PCA could be examined.[46,47] The data are preprocessed before the analysis. As we are interested in between-group differences in correlation structure, the variables are centered in each

*In order to update the PCA solution, there should remain as many variables as components in the blocks.

group, discarding between-group differences in means that may otherwise distort the obtained loadings.[48] Moreover, to focus on the group-specific correlation structures of the variables rather than on their covariances, each variable is rescaled so that its variance equals 1 per group.[48] Thus, preprocessing boils down to standardizing all variables per group. Note that in case the groups contain strongly different numbers of objects, this standardization means that the larger group will dominate the SCA solution that is used as target structure when ranking the variables.

3 | PCA AND SCA

3.1 | Principal component analysis

Principal component analysis[49] reduces the J observed variables to Q components ($Q < J$). To this end, the ith data group $\mathbf{X}_i$ is modeled as follows:

$$\mathbf{X}_i = \mathbf{F}_i\mathbf{B}_i' + \mathbf{E}_i. \tag{1}$$

The score matrix $\mathbf{F}_i(N_i \times Q)$ holds the scores of the N_i objects on the Q components. Note that we restrict the number of components Q to be the same across the groups, because we assume that most of the variables behave similarly. To partially identify the models (partially, because any rotation is still allowed), the variance of each column of $\mathbf{F}_i$ is fixed to 1. Due to the centering of the data blocks, the mean of each column of $\mathbf{F}_i$ equals 0. The matrix $\mathbf{E}_i$ denotes the residual matrix. In principal axes orientation or after orthogonal transformation (see Section 3.4), the loadings in $\mathbf{B}_i$ represent the correlations between the observed variables and the components and allow to label the constructs measured by the components. As Equation 1 suggests, the loading matrices $\mathbf{B}_i(J \times Q)$ are allowed to differ across the groups.

3.2 | Simultaneous component analysis

The multigroup extension of PCA is easily derived from 1 by restricting the loading matrices to be the same across the groups:

$$\mathbf{X}_i = \mathbf{F}_i\mathbf{B}' + \mathbf{E}_i. \tag{2}$$

Here, the variance of each column of the vertically concatenated $\mathbf{F}_i$ matrices is fixed to 1. The obtained model is thus the SCA model with equal pattern restrictions (SCA-P) model.[50] Note that other, more restrictive, SCA variants exist[8] but will not be discussed further in this paper. Again, the loadings in $\mathbf{B}$ can be interpreted as correlations, unless the components are obliquely transformed (see Section 3.4).

3.3 | Deciding on the number of components Q

As mentioned before, the number of components Q is assumed to be the same across the groups in this paper. Sometimes, this number will be decided based on previous studies or a priori hypotheses about the correlation structure of the variables. Alternatively, one can try to assess which number of components yields an optimal balance between fit to the data and model complexity. To do this, we can apply SCA-P with different numbers of components to the data set and conduct a scree test[22,51,52] or parallel analysis.[53]

3.4 | Linear transformation

Once the decision on Q is made, we will have two loading matrices for each group: one based on the SCA-P and one based on the group-specific PCA. We may now linearly transform (or rotate as it is also often called) the obtained group-specific loading structures, to enhance interpretation. Several options exist: each structure may be linearly transformed independently. Alternatively, one may freely transform only one of them and transform the other structure

towards the obtained loadings for the first one to discard arbitrary differences between both structures. Applying such a target linear transformation before investigating the loading patterns' similarity is a common practice in psychological research.[30,37,54,55] In this paper, we follow De Roover et al[26] and use yet another option. Specifically, we first apply a VARIMAX transformation[56] to the SCA-P loadings, followed by an oblique Procrustes transformation (i.e., target transformation) of the group-specific PCA loadings towards the VARIMAX-rotated SCA-P ones. Note that after an oblique transformation, the loadings can have an absolute value higher than 1 and can therefore no longer be interpreted as correlations.

4 | HOW OUTLYINGNESS RANKINGS ARE OBTAINED IN THE EXISTING OUTLYING VARIABLE HEURISTICS

The existing methods to detect outlying variables in column-coupled multigroup data, such as the lower bound congruence method (LBCM)[26] and the lower and resampled upper bound congruence method (LRUBCM),[27] differ in stopping criterion but start from the same outlyingness ranking. This outlyingness ranking is obtained through $J - Q$ iterations. In each iteration, the most outlying variable is determined and removed, and the resulting increase in Tucker's congruence is recorded. Each iteration consists of the following steps:

1. For each corresponding component in the two groups, compute Tucker's congruence between the associated loading columns of the remaining variables in the two groups. Specifically, Tucker's congruence φ_q for the qth component is calculated as follows, with b^i_{jq} indicating the loading of variable j on component q in group i:

$$\varphi_q = \frac{\sum_j b^1_{jq} b^2_{jq}}{\sqrt{\sum_j \left(b^1_{jq}\right)^2 \sum_j \left(b^2_{jq}\right)^2}}. \tag{3}$$

Next, compute the mean of the Q resulting congruence scores: φ^{mean}. For example, for hypothetical data shown in Table 1, the congruence scores for the first and second components amount to 0.9875 and 0.9834, respectively. The mean of these values equals 0.9855.

2. To decide which of the remaining variables is the most outlying, compute variable-specific congruence-after-exclusion scores, quantifying the mean congruence obtained after excluding each variable one by one. The variable with the highest congruence after-exclusion score is considered the most outlying variable.
3. Remove the most outlying variable and re-estimate the group-specific PCA and overall SCA-P solutions. Rotate the group-specific PCA loadings towards the VARIMAX-rotated SCA-P ones. Record the most outlying variable from substep 2 and φ^{mean} value from substep 1.

For the hypothetical example, performing nine iterations yields the outlyingness ranking given in Table 2.

5 | COEFFICIENTS FOR QUANTIFYING THE SIMILARITY OF TWO LOADING STRUCTURES

Tucker's congruence for each component is a common way of quantifying similarities of loading structures.[39,41,38] Yet many alternatives have been proposed, both in psychology and in other disciplines: applying Tucker's coefficient to all loadings at once or per variable rather than per component,[28,54] computing the Rv coefficient[28,32,33] or a modified version thereof,[34] focusing on the angle between subspaces,[35,36] and calculating the RMSD between variable-specific loading, component-specific loading, or all loadings.[37,38] In this section, we will review these measures, focusing on four characteristics: which loadings are included (per variable, per component, all), which types of transformation are they affected by, which range of values can they take, and do they measure similarity or dissimilarity. Table 3 presents an overview of these characteristics for the different alternatives.

TABLE 1 Group-specific true loading matrices and the ones recovered by principal component analysis (PCA) for a simulated data set

	B_1		B_2	
True loading matrices	**0.77**	0	**0.77**	0
	0.77	0	**0.77**	0
	0.77	0	**0.77**	0
	0.77	0	**0.77**	0
	0.77	0	**0.77**	0
	0	**0.77**	0	**0.77**
	0	**0.77**	0	**0.77**
	0	**0.77**	0	**0.77**
	0	**0.77**	0	**0.77**
	0	**0.77**	0	**0.77**
	0.77	0	0.39	**0.67**
PCA loading matrices	**0.70**	0.07	**0.87**	0.06
	0.71	−0.05	**0.86**	0.05
	0.71	0.16	**0.83**	0.13
	0.75	0.07	**0.88**	0.03
	0.68	0.12	**0.86**	0.02
	0.00	**0.75**	0.14	**0.60**
	0.03	**0.81**	0.12	**0.71**
	0.12	**0.79**	−0.05	**0.61**
	0.16	**0.80**	0.10	**0.67**
	0.11	**0.77**	0.05	**0.59**
	0.65	0.21	**0.63**	**0.40**

Note. The data set was simulated with two components, consisting of five nonoutlying variables each, one outlying variable with very high degree of outlyingness, 40% of error, and 250 observations in each group. Each group-specific loading matrix was obliquely transformed towards the VARIMAX-transformed overall structure (obtained by simultaneous component analysis [SCA]). The loadings with an absolute value higher than .40 are printed in bold.

5.1 | Tucker's congruence coefficient

Equation 3 showed the formula for computing component-specific congruence scores. This formula can easily be adapted to obtain a variable-specific version:

$$\varphi_j = \frac{\sum_q b_{jq}^1 b_{jq}^2}{\sqrt{\sum_q \left(b_{jq}^1\right)^2 \sum_q \left(b_{jq}^2\right)^2}}, \tag{4}$$

or a version for all loadings at once,[28] which we call the full version:

$$\varphi_{full} = \frac{\sum_j \sum_q b_{jq}^1 b_{jq}^2}{\sqrt{\sum_j \sum_q \left(b_{jq}^1\right)^2 \sum_j \sum_q \left(b_{jq}^2\right)^2}}. \tag{5}$$

For the hypothetical data in Table 1, the variable-specific congruence scores range from 0.9996 to 0.9674. To obtain one overall congruence score, we compute the mean of the variable-specific scores, which amounts to 0.9890. The full

TABLE 2 Outlyingness rankings for the toy example for the different similarity measures

Measure	Mean statistics	Most outlying variable	Measure	Mean statistics	Most outlying variable	Measure	Mean statistics	Most outlying variable
φ_q	0.9855	11	φ_j	0.9890	11	φ_{full}	0.9715	8
	0.9913	8		0.9915	8		0.9768	10
	0.9935	5		0.9942	6		0.9812	9
	0.9940	6		0.9961	7		0.9831	5
	0.9962	7		0.9966	2		0.9878	2
	0.9970	2		0.9981	1		0.9886	11
	0.9983	1		0.9996	9		0.9951	7
	0.9995	3		0.9977	3		0.9975	6
	0.9945	9		0.9995	4		0.9431	4
RMSD_q	0.1276	8	RMSD_j	0.1251	8	$\text{RMSD}_{\text{full}}$	0.1276	8
	0.1186	5		0.1167	5		0.1187	5
	0.1111	10		0.1097	10		0.1111	10
	0.1032	9		0.1026	9		0.1037	9
	0.0955	2		0.0918	2		0.0957	2
	0.0901	11		0.0871	11		0.0901	11
	0.0604	1		0.0606	1		0.0618	1
	0.0465	4		0.0462	7		0.0468	7
	0.0385	3		0.0323	6		0.0362	6
Rv_{unr}	0.9186	6	Rv_{obl}	0.9248	8	$Rv_{\text{obl}}^{\text{mod}}$	0.9232	8
	0.9358	7		0.9445	10		0.9445	10
	0.9561	8		0.9600	9		0.9610	9
	0.9772	9		0.9735	6		0.9749	6
	0.9855	10		0.9837	2		0.9831	2
	0.9724	5		0.9841	7		0.9857	11
	0.9794	11		0.9619	5		0.9931	7
	0.9566	1		0.9927	11		0.9885	4
	0.9534	2		0.9752	4		0.9995	5
$Rv_{\text{unr}}^{\text{mod}}$	0.9158	6	$\cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2}$	0.9771	11			
	0.9357	7		0.9880	6			
	0.9598	8		0.9931	8			
	0.9819	10		0.9938	5			
	0.9919	9		0.9956	7			
	0.9668	5		0.9948	2			
	0.9769	11		0.9977	1			
	0.9469	2		0.9993	3			
	0.9463	1		0.9880	10			

Note. The loadings of the hypothetical example are shown in Table 1.

congruence coefficient in Equation 5 evaluates the similarity of all loadings at once and thus does not require any further aggregation steps. It equals 0.9715 for the hypothetical example.

All three congruence coefficients measure the proportional similarity of the considered loadings and are thus insensitive to differences in scaling. Possible values range from −1 (loadings are perfectly proportional and have opposite signs) to 1 (loadings are perfectly proportional and have the same sign). A value close to 0 means that the vectors are

TABLE 3 Characteristics of the different similarity indices

Similarity measure	Which loadings included			Sensitive to which transformations	Similarity or dissimilarity	Range
	Variable	Component	All			
Congruence	✓	✓	✓	All	Similarity	[−1, 1]
Rv	×	×	✓	Oblique	Similarity	[0, 1]
Modified *Rv*	×	×	✓	Oblique	Similarity	[0, 1]
The cosine of an angle between subspaces	×	×	✓	None	Similarity	[0, 1]
RMSD	✓	✓	✓	All	Dissimilarity	≥0

very dissimilar, whereas absolute values close to 1 indicate strong similarity. Because all congruence coefficients are based on specific loading values, which are affected by transformations, they are all transformation dependent.

5.2 | The *Rv* coefficient

The *Rv* coefficient is a similarity coefficient for positive semidefinite matrices, which are thus square and symmetric.[28,32,33] Generally, loading matrices are rectangular, however, because the number of components Q is lower than the number of variables J. Therefore, the *Rv* coefficient is computed on the matrices that result from multiplying the loading matrices by their transpose, thus on $\mathbf{C}_1\left(c_{ij}^1\right) = \mathbf{B}_1\mathbf{B}_1'$ and $\mathbf{C}_2\left(c_{ij}^2\right) = \mathbf{B}_2\mathbf{B}_2'$, where $i,j = 1,\ldots J$:

$$Rv = \frac{\sum_i\sum_j c_{ij}^1 c_{ij}^2}{\sqrt{\sum_i\sum_j\left(c_{ij}^1\right)^2\sum_i\sum_j\left(c_{ij}^2\right)^2}}. \tag{6}$$

Comparing Equations 5 and 6 shows that the computation of the *Rv* coefficient boils down to computing the full congruence of $\mathbf{B}_1\mathbf{B}_1'$ and $\mathbf{B}_2\mathbf{B}_2'$. On the basis of the Cauchy-Schwartz inequality, the *Rv* coefficient ranges from 0 to 1 and can also be considered a similarity measure. In comparison with the congruence coefficient, the *Rv* coefficient is completely unaffected by orthogonal transformations. Indeed, if $\mathbf{T}$ is an orthogonal transformation matrix, $(\mathbf{B}_1\mathbf{T})(\mathbf{B}_1\mathbf{T})' = \mathbf{B}_1\mathbf{T}\mathbf{T}'\mathbf{B}_1' = \mathbf{B}_1\mathbf{B}_1'$. However, *Rv* values are sensitive to oblique transformations. Nevertheless, in comparison with all the other coefficients that are sensitive towards the oblique transformation, *Rv* (as well as modified *Rv*) is less affected by such transformations, because it is based on the derived matrices, $\mathbf{C}_1$ and $\mathbf{C}_2$, instead of being based directly on the loadings. Therefore, we will consider two *Rv* versions in this paper: *Rv* on the untransformed (unrotated) loadings (Rv_{unr}) and *Rv* on the obliquely transformed ones (Rv_{obl}), as this will show us how much the performance results depend on the decision to apply oblique rather than orthogonal transformations. For the obliquely transformed loadings shown in Table 1, the Rv_{obl} coefficient equals 0.9248.

5.3 | The modified *Rv* coefficient

Smilde et al[34] showed that, although one would expect *Rv* to be close to 0 for random matrices, it is artificially high for matrices with a small number of rows. Because this bias can be traced back to the diagonal of the $\mathbf{B}_1\mathbf{B}_1'$ and $\mathbf{B}_2\mathbf{B}_2'$ matrices, they suggested to replace the diagonal elements with zeros, yielding the modified *Rv* coefficient Rv^{mod}. This coefficient has the same characteristics as the *Rv* coefficient in that it is a similarity measure, ranges from 0 to 1, and is affected by oblique transformations only. Thus, we will again consider two versions: $Rv_{\text{unr}}^{\text{mod}}$ and $Rv_{\text{obl}}^{\text{mod}}$. For the loadings shown in Table 1, $Rv_{\text{obl}}^{\text{mod}}$ hardly differs from Rv_{obl}: 0.9232.

5.4 | Cosine of the angle between the two subspaces

Another interesting coefficient boils down to calculating the cosine of the angle between the two subspaces spanned by the two loading matrices.[35,36] It is defined as follows:

$$cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2} = cos\vartheta_1 cos\vartheta_2 \cdots cos\vartheta_q, \tag{7}$$

where $\vartheta_1,\vartheta_2...,\vartheta_q$ are the principal angles, that is, the angles between the corresponding principal vectors that span the $\mathbf{B}_1$ and $\mathbf{B}_2$ spaces. By definition, principal angles range between 0 and $\pi/2$, and therefore, the cosine of the angle between the subspaces ranges from 0 to 1. In this paper, we use the MATLAB implementation to calculate the cosine of the angle between two subspaces. This implementation is based on numerical methods elaborated by Björck and Golub[35] and Wedin.[57] Values close to 0 indicate that loading patterns are dissimilar, whereas values close to 1 indicate strong similarity. The cosine of an angle between the subspaces is not affected by transformations, as no transformation can change the subspace that is spanned by the loading matrix. For the loadings in Table 1, the cosine of the angle between the subspaces equals 0.9771.

5.5 | RMSD of the loadings

Finally, one can compute the RMSD[37,38] of component-specific loadings:

$$\mathrm{RMSD}_q = \sqrt{\frac{\sum_j \left(b_{jq}^1 - b_{jq}^2\right)^2}{J}}, \tag{8}$$

variable-specific loadings:

$$\mathrm{RMSD}_j = \sqrt{\frac{\sum_q \left(b_{jq}^1 - b_{jq}^2\right)^2}{Q}}, \tag{9}$$

or all loadings at once:

$$\mathrm{RMSD}_{\mathrm{full}} = \sqrt{\frac{\sum_{\mathrm{j}}\sum_{\mathrm{q}} \left(\mathrm{b}_{\mathrm{jq}}^1 - \mathrm{b}_{\mathrm{jq}}^2\right)^2}{\mathrm{JQ}}}. \tag{10}$$

As was the case for the congruence coefficients, the first two RMSD measures have to be averaged across the components and the variables, respectively, to obtain an overall value. For example, for the loadings in Table 1, the mean RMSD_q amounts to 0.1276, and the mean RMSD_j to 0.1251; $\mathrm{RMSD}_{\mathrm{full}}$ equals 0.1276. RMSD values close to 0 indicate strong similarity, whereas higher values suggest dissimilarity. Thus, unlike all the other measures discussed in this paper, the RMSD is a dissimilarity measure. Because absolute differences for specific loadings constitute the basic building block of the RMSD, it is sensitive to every type of transformation.

5.6 | Comparison of the different coefficients and possible effects on the outlyingness ranking

Next to the characteristics discussed above, a further investigation of the different coefficients reveals some interesting differences that might affect the outlyingness ranking. First, unlike the measures that only assess proportional similarity, the RMSD coefficients are stricter and quantify absolute similarity of the loadings. For example, let us assume that a variable has the following two loading patterns in the two groups: [0.4, 0.5, 0.4] and [0.6, 0.75, 0.6]. These patterns are perfectly proportional as the loadings for the second group are 1.5 times larger than those in the first group. So φ_j would equal 1. RMSD_j would be larger than 0, indicating some dissimilarity, however, because the loadings are clearly larger in the second group. As a result, replacing the congruence measure by an RMSD coefficient might lead to a different variable being labeled as the most outlying one and, in general, to a different outlyingness ranking. Indeed,

Table 2 shows the outlyingness rankings for the toy example for the different similarity measures, and it demonstrates that $RMSD_q$ results in the eighth variable being the most outlying one, whereas component-specific congruence φ_q pinpoints the 11th variable as most outlying.

Second, we also expect that the subset of loadings that is considered by the coefficient can determine what kind of differences they are prone to detect. In particular, component-specific Tucker's congruence scores φ_q (respectively, variable-specific scores φ_j) assess the proportional similarity of each separate component (respectively, variable), while including all loadings yields a coefficient φ_{full} that is more sensitive to the proportional similarity of the whole configuration. Indeed, if all components are proportionally similar but the exact proportions change across the components, this will yield very high φ_q scores but a lower φ_{full} score, as here, the proportion should be the same across the components to obtain a high score. As expected, for the toy example, φ_{full} is lower than φ_j and φ_q (Table 2).

Third, we also expect sensitivity differences between the coefficients that are applied at the full level, thus between φ_{full}, Rv, and $\cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2}$. Specifically, the Rv coefficient boils down to computing φ_{full} on different matrices, that is, the loading matrices multiplied by their transpose. In these new matrices, loading differences might be less clear and harder to detect because multiple loadings are combined in each entry.

6 | SIMULATION STUDY

6.1 | Problem

In this section, we present a simulation study in which we evaluate whether and how the different similarity indices affect the obtained outlyingness ranking. Moreover, we explore how the performance is further influenced by the following seven factors: (1) the degree of outlyingness, (2) the error level e, (3) the number of components, (4) the ratio of nonoutlying variables to components, (5) the number of outlying variables, (6) the number of observations per group, and (7) the component variances and correlations. In general, we expect that the outlyingness ranking will be less correct for lower degrees of outlyingness, more error, more components, less nonoutlying variables per component, more outlying variables, less observations per group, and larger differences in component variances and correlations. Nevertheless, some of the similarity coefficients might be more robust towards the harder data configurations than others. Finally, we will also briefly examine whether performance is affected by differences in the number of observations per group. We expect that such differences will impact performance in two interacting ways: first, the smaller the size of the smaller group, the worse the performance. Second, the larger the average group size, the better the performance. So results will be good if both groups are sufficiently large and especially bad if both groups are too small. These expectations are based on the well-known fact that the recovery of PCA loadings depends on the number of observations.

Concerning the manipulation of the component variances and correlations, note that in case of component variances that differ from 1, the identification restrictions on the component variances will result in a component-specific rescaling of the estimated loadings, which will therefore deviate from the "true" loadings. This effect is only partly counteracted by standardizing the data per group. As a result, the measures that are more robust towards component-specific disturbances in the loading structures, such as φ_q and $RMSD_q$, should be less affected by this manipulation.

6.2 | Design and procedure

In the main simulation, each simulated data set consisted of two equally sized groups; the results when varying group size within the data sets will be discussed in Section 6.3. The following seven "between" factors were systematically varied in a complete factorial design:

1. The degree of outlyingness: very high, high, medium, and low.
2. The error level e, which is the expected proportion of error variance in the data groups: 0.20 and 0.40.
3. The number of components: 2, 3, and 4.
4. The ratio of nonoutlying variables to components: 3, 5, and 7.

5. The number of outlying variables: 1, 2, 3, and 4.
6. The number of observations per group: 100, 250, and 750.
7. The component variances and correlations: for the first level, all the component variances are set equal to 1, and all the component correlations to 0. For the second level, the component variances were uniformly sampled between 0.25 and 1.75, and the components were all uncorrelated. For the third level, the component variances and correlations were uniformly sampled between 0.25 and 1.75 and between −0.5 and 0.5, respectively.

The factorial design thus contains $4 \times 2 \times 3 \times 3 \times 4 \times 3 \times 3 = 2{,}592$ cells. For each cell of this design, 100 data matrices $\mathbf{X}$ were generated, yielding 259,200 data sets. Each data matrix consists of two groups. For each group, a component score matrix $\mathbf{F}_i$ was randomly sampled from a multivariate normal distribution with the component variances and correlations determined by design factor 7.

To generate the two group-specific loading matrices, $\mathbf{B}_i$, the following steps were performed. To each of the components, three, five, or seven different nonoutlying variables were assigned according to design factor 4, implying that those variables have a loading of one on that component and zero loadings on the other components. In Group 1, outlying variables also had a loading of one on a single component, whereas in the other group, they received not only a loading b_{outl1} on the corresponding component but also a loading b_{outl2} on another component in the following manner. The b_{outl1} loadings of the outlying variables were spread as evenly as possible across the components. For example, in case of three outlying variables and three components, each of those components receives one outlying variable only. As another example, when there are four outlying variables and three components, one randomly selected component contains two outlying variables, and the other two components only one. The components that receive the b_{outl2} loadings were selected in a similar way. The values of b_{outl1} and b_{outl2} depend on the degree of outlyingness. For the very high, high, medium, and low degrees of outlyingness, the combinations of b_{outl1} and b_{outl2} are $\sqrt{0.25}$ and $\sqrt{0.75}$, $\sqrt{0.50}$ and $\sqrt{0.50}$, $\sqrt{0.75}$ and $\sqrt{0.25}$, and $\sqrt{0.85}$ and $\sqrt{0.15}$, respectively. Afterward, the loading matrices were rescaled by multiplying them to $\sqrt{1-e}$, to take the error level (design factor 2) into account. For instance, the loadings in Table 1, which pertain to a condition with five nonoutlying variables per component and very high outlyingness, were multiplied by $\sqrt{1-0.4}$ because the error level equals 0.4.

For both groups, the error matrices, $\mathbf{E}_i$, were randomly sampled from a multivariate normal distribution with zero means and an identity covariance matrix and were rescaled by multiplying it with $\sqrt{e}$ to obtain the desired error level (design factor 2). Finally, each $\mathbf{X}_i$ is computed as $\mathbf{F}_i\mathbf{B}_i' + \mathbf{E}_i$.

When analyzing the simulated data, we obtained 11 outlyingness rankings by using the following 11 similarity measures and transformation choices: φ_q, φ_j, φ_{full}, Rv_{obl}, Rv_{unr}, $Rv_{\text{unr}}^{\text{mod}} Rv_{\text{obl}}^{\text{mod}}$, $\cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2}$, RMSD_q, RMSD_j, and $\text{RMSD}_{\text{full}}$. Note that when using φ and RMSD, loadings are always obliquely transformed using Procrustes transformation.

The correctness of the "outlyingness ranking" was used as a performance measure. We knew beforehand how many outlying variables each data set contained, and we used this knowledge when analyzing the data. For example, in case the data were simulated with four outlying variables, we consider the outlyingness ranking to be correct when the first four variables in the ranking are the truly outlying ones.

An attentive reader might wonder whether and how differences in group size would influence the obtained outlyingness ranking. In order to not make the simulation design too complex, we decided to check for this after we had identified the best performing measure. We expected that the following two factors will have an effect on the performance: the size of the smallest group and the average size of the groups. To inspect this, we fixed the other six "between" factors as follows: the degree of outlyingness to low, the error level to 40%, the number of components to 3, the ratio of nonoutlying variables to components to 7, the number of outlying variables to 4, and the component variances and correlations to 1 and 0, respectively. With respect to the group sizes, we used the following approach: we varied the average group size from 100 to 500 in nine equal steps, yielding 100, 150, 200, and so on. We also varied the difference between the group sizes from 0 to a maximum 800 in equal steps. For each level of average group sizes, we started with a minimum group size of 100 for the smallest group and increased this size with 50 until it reached the desired average group size. As a result, the number of possible combinations depends on the average group size. Specifically, for the average sample size of 100, there is only one combination, [100 100]; for 150, there are two combinations, [100 200] and [150 150]; and for 500, there are nine different combinations ranging from [100 900], [150 850], ..., until [500 500]. Table 4 shows all the combinations of the group sizes that were simulated. As results might be somewhat more unstable across replications due to the group size differences, we generated 1,000 data sets for each of these combinations and analyzed them using Tucker's congruence, φ_q. Again, the correctness of the outlyingness ranking was used as a performance measure.

TABLE 4 The number of data sets with correct outlyingness rankings (depicted in black) when varying the group sizes within the data

	Difference in group size								
Average group size	800	700	600	500	400	300	200	100	0
100									100 100
									347
150								100 200	150 150
								528	607
200							100 300	150 250	200 200
							627	607	824
250						100 400	150 350	200 300	250 250
						664	848	905	935
300					100 500	150 450	200 400	250 350	300 300
					685	897	953	965	973
350				100 600	150 550	200 500	250 450	300 400	350 350
				705	907	973	987	992	990
400			100 700	150 650	200 600	250 550	300 500	350 450	400 400
			719	907	980	993	995	998	996
450		100 800	150 750	200 700	250 650	300 600	350 550	400 500	450 450
		704	914	979	990	997	1,000	999	998
500	100 900	150 850	200 800	250 750	300 700	350 650	400 600	450 550	500 500
	726	927	987	990	997	999	999	1,000	1,000

Note. Each data set was simulated with low outlyingness, 40% of error level, three components, seven nonoutlying variables assigned to each component, four outlying variables, and one and zero component variances and correlations, respectively. Both group sizes (depicted in gray) depend on the average group size and on the difference between the two group sizes.

In order to investigate whether it matters which of both data blocks is largest, we checked two different scenarios. In the first scenario, the bigger group showed perfect simple structure, whereas the smaller group was characterized by cross loadings. In the second scenario, the loadings were generated the other way around so with the larger group showing cross loadings.

6.3 | Results

The overall performance is a good place to start comparing the 11 ways of extracting outlyingness rankings. The top plot of Figure 1 shows the percentage of simulated data sets for which the outlying variables are ranked correctly. This plot reveals that φ_q shows the best overall performance. To check whether this measure is consistently the best choice across the different "between" factor design cells, we first ranked the 11 measures in each of the 2,592 cells according to performance: the measures yielding the highest number of correct outlying rankings (note that more than one measure can yield the same number of correct outlying rankings) receiving the lowest rank (i.e., 1), the measures yielding the second highest number of correct outlying rankings received rank number 2, and so on. Afterward, we calculated the average rank across the cells per measure and visualized these averages in the bottom plot of Figure 1. Both performance orderings of the 11 measures are quite comparable, indicating that φ_q performs very well in all cells of the design. The other two congruence measures, φ_j and φ_{full}, take the second and third places, followed by the three RMSD-based measures and $\cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2}$. The four *Rv*-based measures perform the worst, although the ones considering oblique transformations do better than the ones without transformation.

This conclusion is further strengthened by Figure 2. In this figure, the performance of φ_q is marked with bigger dots and that of the other measures with smaller dots, whereas the *X*-axis reflects the different between-cells of the design. φ_q is clearly the best measure in the biggest number of cells but was outperformed by other measures in 194 out of the 2,592 cells. These cells, which show no coherent design factor pattern and thus seem to be sampled rather randomly

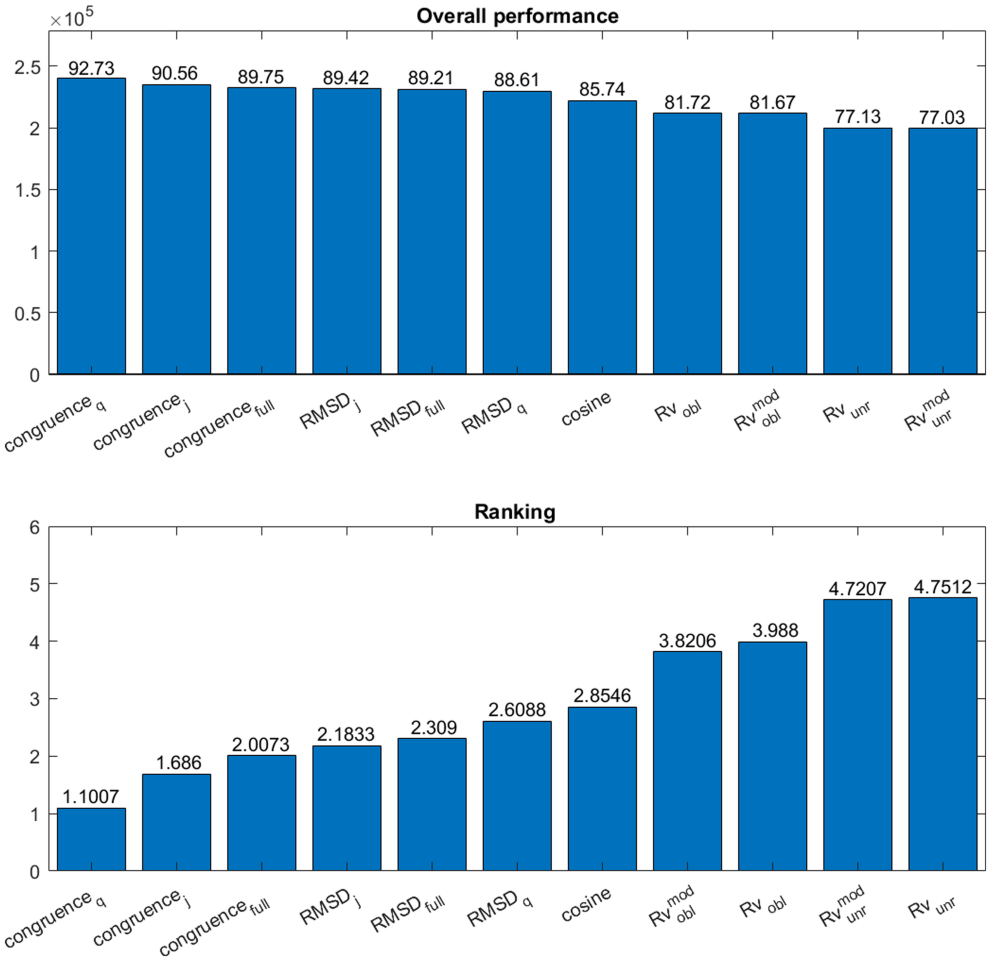


FIGURE 1 The top plot shows the overall percentage of data sets with correct outlyingness rankings for the different similarity measures. The bottom plot shows the mean ranking for each of the measures. The rank for each measure in each cell of the simulation design was determined in the following manner: the measures yielding the highest number of correct outlyingness rankings (i.e., the highest number of correctly identified data sets out of 100) received rank 1—if there was more than one measure with the highest number, all of them were ranked as 1; the measures with the second highest number were ranked second and so on. Finally, for each measure, the average rank across the cells of the design was calculated. For both plots, the measures are ordered from better to worst performing

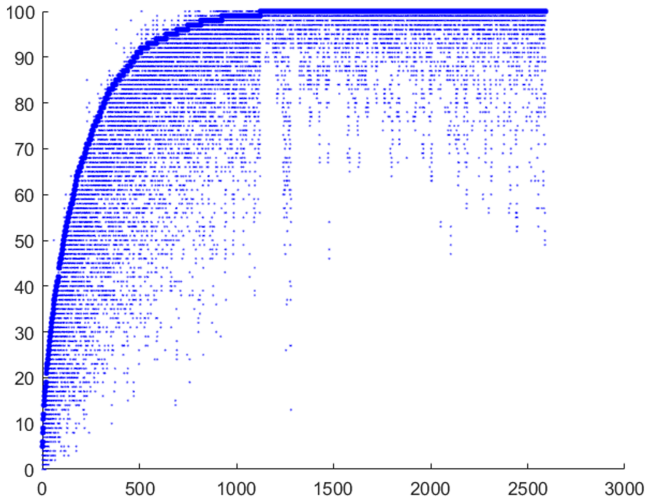


FIGURE 2 The plot of the number of correct outlyingness rankings for the 2,592 design cells across different measures. The performance of φ_q is marked with bigger dots, and all the other measures are marked in smaller dots. The *X*-axis reflects the different between-cells of the design, and the cells are ordered according to the increasing performance of φ_q

from the design, correspond to the spikes that cross the line of large dots in Figure 2. We inspected which measures do better than φ_q in these cells, noting that multiple measures can yield better results than φ_q in a specific cell. Figure 3 shows that all other measures yield better results in specific cells. $\cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2}$ is the clear winner in this regard because it outperforms φ_q in 83 cells, whereas the *Rv*-based measures only beat φ_q in a low number of cells. Also, here, not transforming the loadings yields the worst *Rv* results.

Furthermore, we only consider the outlyingness ranking to be correct when all the outlying variables are ranked correctly, thus if the n outlying variables take the top n positions in the outlyingness ranking. This implies that even if the complete ranking is considered incorrect, some of the outlying variables may still be among the top n variables and

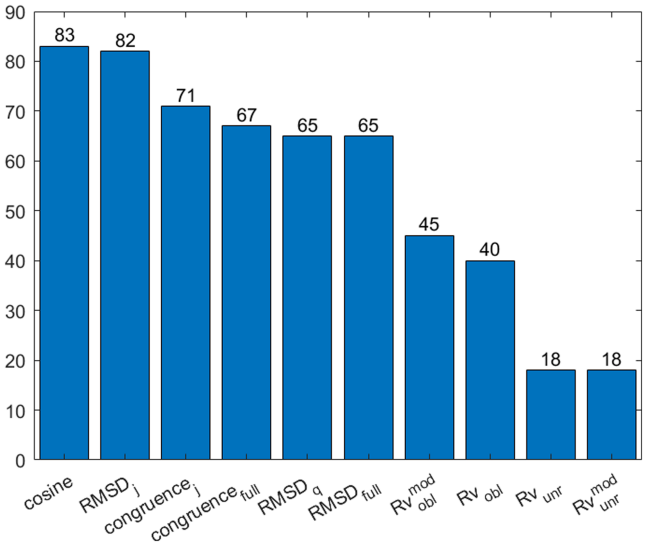


FIGURE 3 The bar graphs of the better performing measures in 194 cells of the simulation design where φ_q was not ranked as the best combination

are thus found. To investigate this further, we looked at the incorrect outlyingness rankings that were produced by φ_q for 18,847 out of the 259,200 data sets. From these incorrectly ranked data sets, 2,558 data sets contained only one outlying variable (which were all incorrectly ranked), 4,195 data sets contained two outlying variables, yielding 8,390 outlying variables of which 4,654 were not found, 5,541 data sets contained three outlying variables, yielding 16,623 outlying

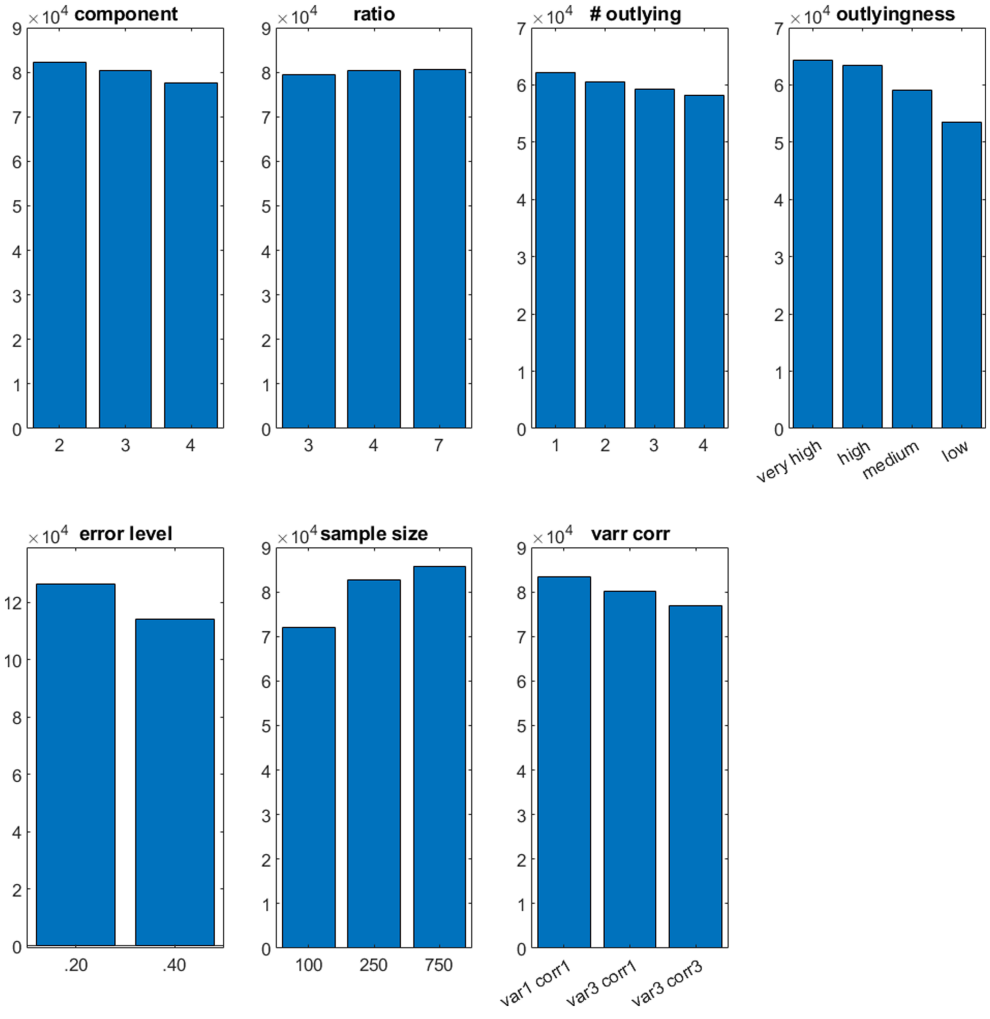


FIGURE 4 Performance of φ_q as a function of the seven manipulated factors

TABLE 5 Group-specific principal component analysis (PCA) loadings for the bread data, where the first group consists of the first, second, and seventh panelists and the second group comprises the rest of the panelists

	B_1		B_2		Overall B-VARIMAX	
Attributes	**Salt**	**Yeast and disfavor**	**Salt**	**Yeast**	**Salt**	**Yeast**
Bread-od	−0.20	−0.20	**0.50**	0.17	0.34	0.08
Yeast-od	−0.04	**0.87**	0.14	**0.69**	0.09	**0.75**
Off-flav	−0.32	**−0.43**	**−0.51**	**−0.51**	**−0.49**	**−0.52**
Color	**0.38**	−0.18	0.34	0.43	**0.43**	0.09
Moisture	**0.80**	0.26	**0.74**	0.21	**0.77**	0.22
Tough	**0.78**	−0.35	**0.85**	−0.08	**0.83**	−0.23
Salt-t	**0.76**	−0.41	**0.82**	−0.32	**0.80**	−0.36
Sweet-t	**−0.80**	−0.10	**−0.81**	−0.17	**−0.81**	−0.16
Yeast-t	0.03	**0.84**	−0.09	**0.75**	−0.05	**0.79**
Other-t	**−0.61**	**−0.68**	**−0.67**	−0.26	**−0.63**	**−0.47**
Total	**−0.69**	0.03	**0.75**	0.01	0.31	0.17

Note. Group-specific solutions are obliquely transformed towards the VARIMAX simultaneous component analysis (SCA) loadings. The loadings with an absolute value higher than .40 are printed in bold.

TABLE 6 Outlyingness rankings for the loading differences across the two clusters of panelists

Measure	**Mean statistics**	**Most outlying variable**	**Measure**	**Mean statistics**	**Most outlying variable**	**Measure**	**Mean statistics**	**Most outlying variable**
φ_q	0.7425	Total	φ_j	0.5540	Total	φ_{full}	0.7133	Total
	0.8871	Bread-od		0.7431	Bread-od		0.8907	Bread-od
	0.9319	Color		0.9071	Color		0.9384	Color
	0.9758	Other-t		0.9822	Other-t		0.9731	Other-t
	0.9857	Tough		0.9875	Yeast-od		0.9829	Tough
	0.9914	Yeast-od		0.9942	Tough		0.9882	Yeast-od
	0.9968	Off-flav		0.9960	Off-flav		0.9953	Off-flav
	0.9959	Salt-t		0.9977	Salt-t		0.9972	Salt-t
	0.9920	Sweet-t		0.9972	Moisture		0.9974	Sweet-t
RMSD_q	0.3835	Total	RMSD_j	0.2847	Total	$\text{RMSD}_{\text{full}}$	0.3983	Total
	0.2472	Bread-od		0.1890	Bread-od		0.2491	Bread-od
	0.1790	Color		0.1500	Color		0.1972	Color
	0.1337	Other-t		0.1204	Other-t		0.1395	Other-t
	0.1094	Yeast-od		0.0898	Yeast-od		0.1095	Tough
	0.0779	Tough		0.0621	Tough		0.0966	Yeast-od
	0.0616	Moisture		0.0579	Moisture		0.0621	Moisture
	0.0630	Sweet-t		0.0616	Sweet-t		0.0640	Sweet-t
	0.0432	Salt-t		0.0437	Salt-t		0.0446	Salt-t
Rv_{unr}	0.4752	Total	Rv_{obl}	0.4820	Total	$Rv_{\text{obl}}^{\text{mod}}$	0.3751	Total
	0.7765	Bread-od		0.7973	Bread-od		0.7579	Bread-od
	0.8750	Yeast-od		0.9004	Color		0.8843	Color
	0.9227	Yest-t		0.9441	Other-t		0.9376	Other-t
	0.8805	Color		0.9614	Tough		0.9547	Yeast-od
	0.9629	Tough		0.9715	Yeast-od		0.9815	Tough

(Continues)

TABLE 6 (Continued)

Measure	Mean statistics	Most outlying variable	Measure	Mean statistics	Most outlying variable	Measure	Mean statistics	Most outlying variable
	0.9686	Off-flav		0.9886	Off-flav		0.9848	Yest-t
	0.9764	Other-t		0.9943	Yest-t		0.9812	Off-flav
	0.9762	Moisture		0.9779	Sweet-t		0.9610	Moisture
$Rv_{\text{unr}}^{\text{mod}}$	0.3696	Total	$\cos\vartheta_{\mathbf{B}_1,\mathbf{B}_2}$	0.6628	Total			
	0.7330	Bread-od		0.8385	Color			
	0.8501	Yeast-od		0.8885	Bread-od			
	0.9069	Yest-t		0.9629	Other-t			
	0.8514	Color		0.9816	Tough			
	0.9614	Tough		0.9875	Yeast-od			
	0.9771	Salt-t		0.9962	Moisture			
	0.9498	Moisture		0.9937	Off-flav			
	0.9785	Other-t		0.9819	Yest-t			

variables of which 6,838 remained unfound, and 6,553 data sets with four outlying variables, yielding 26,212 outlying variables of which 8,758 were incorrectly ranked.

Finally, the influence of the different "between" factors on the performance of φ_q can be observed from Figure 4. As expected, increasing the number of components, number of outlying variables, error level, differences in component variances and correlations and decreasing the number of observations, the nonoutlying variables-to-components ratio, and the level of outlyingness make it harder to rank outlying variables correctly. However, some of these factors influence performance less than others.

Table 4 shows the number of data sets with correct variable rankings in each design cell, for the scenario in which the smaller groups showed simple structure. We observe that when the smallest group size is close to 100, increasing the sample size of the second group has a strong influence. When the small group is already quite large, the average group size matters more. The results for the alternative scenario (i.e., larger groups showed simple structure) were very comparable with the largest absolute cell difference amounting to −26 and the mean cell difference amounting to 7.6.

6.3.1 | Conclusion

The main objective of the simulation study was to find the best performing measure and to check whether this measure is consistently the best across all the cells of the design. The results showed that φ_q performs better than the other 10 measures investigated in this paper, with this result being hardly affected by the data characteristics manipulated in this study. The *Rv*-based measures yield the worst results, which might be explained by the third discussed difference in Section 5.6. This implies that if an *Rv*-based measure yields a different outlyingness ranking than φ_q, the *Rv*-based one is probably less trustworthy.

7 | ILLUSTRATIVE APPLICATION

In this section, we reanalyze the bread data that were reported on before[21,23] discussing CP and clusterwise SCA analyses, respectively. The data can be downloaded from http://www.models.kvl.dk/Sensory_Bread. Two replicates of five different breads were baked, resulting in 10 different samples. These bread samples were evaluated on 11 attributes by eight trained panelists (i.e., assessors) in a fixed vocabulary profiling analysis. We follow up on the analysis by De Roover et al.[21] These authors investigated whether the equal components assumption that underlies the CP analysis actually holds up in the bread data. Specifically, they investigated whether the attributes were used in different ways by the panelists. To this end, they considered the samples by attributes data of each panelist as a different group and applied clusterwise SCA. This technique partitions the groups into mutually exclusive clusters, according to the

correlation structure of the variables, and fits an SCA model to the data within each cluster. The analysis yielded two clusters of panelists and associated components. Albeit interesting, this analysis sheds no direct light on which variables are used most differently in the two clusters. To answer this question, we will extract the outlyingness ranking of the attributes using the different measures studied in this paper.

We split the data into two groups that correspond to the clusters of panelists identified by De Roover et al.[21] So the first data group is the vertical concatenation of the data of the panelists that belong to their first cluster, whereas the second data group compiles the data of the panelists in the second cluster. Like De Roover et al,[21] we preprocessed the data by standardizing the attributes of each panelist and extracted two components. Table 5 shows the VARIMAX SCA loadings (across all data) and the obliquely transformed cluster-specific loadings. The first SCA component is characterized by high-positive loadings of "color," "moisture," "tough," and "salt-t," whereas "off flavor," "sweet taste," and "other taste" have lower loadings. This loading pattern makes sense as salt on the one hand gives a tougher structure to bread, causes moisture retention, and on the other hand masks bad flavors. Because of the central role of "salt" in this interpretation, we labeled this component as "salt." We labeled the second component as "yeast" due to the high loadings of "yeast odor" and "yeast taste." "Off flavor" and "other taste" have very strong negative loadings on this component, which might indicate that when yeast has such a dominant taste, other tastes can become hard to notice. The group-specific (or in this case, cluster-specific) loadings can in general be interpreted in a similar manner. In particular, the attributes that are related to saltiness load highly on the first component, and yeast-related attributes load highly on the second component. However, there are subtle loading differences for some other attributes between the groups.

Table 6 shows the outlyingness rankings obtained with the different measures. The list of the four most outlying variables that is most consistent across the measures consists of "total," "bread odor," "color," and "other taste." The Rv-based results depend on whether or not the loadings are obliquely transformed, whereas Rv_{obl} and $Rv_{\text{obl}}^{\text{mod}}$ also yield these same four outlying variables. Rv_{unr} and $Rv_{\text{unr}}^{\text{mod}}$ produce a somewhat different list: "total," "bread odor," "yeast odor," and "yeast taste."

As discussed in Sections 5.2 and 5.3, these measures are affected by oblique transformation, which explains these differences between the four outlyingness ranking lists provided by the Rv-based measures. Moreover, simulations showed (Section 5.6) that the two Rv measures that are computed on the original loadings had a clearly inferior performance in comparison with the other measures; thus, our finding suggests that the data contain two strongly outlying variables, "total" and "bread odor," and possibly two weaker ones, "color" and "other taste."

7.1 | Conclusion

According to Bro,[23] the panelists should have very similar responses, if they are well trained, so that averaging responses over panelists should be reasonable. The analysis of the bread sensory data showed that there are differences between the panelists, however. In particular, panelists seem to use two attributes ("total" and "bread odor") in a different way. These findings raise the question whether it is appropriate to apply simultaneous analyses to the data as it is. It might make sense to remove these outlying attributes before running further analyses on these data.

8 | DISCUSSION

This paper focused on methods to check whether components can be considered the same across the groups of a multivariate multigroup data set and to find the outlying variables that cause possible discrepancies. This research line originated from psychometrics. Overall, this paper had two main aims. The first aim was to introduce this line of research to chemometrics and sensory profiling. We think that looking for possible outlying variables can lead to crucial information and insight into data as well as suggest some steps that need to be taken to make sure that the assumptions made by some multiway models hold. Indeed, CP, one of the most popular multiway data analysis techniques in chemometrics, implies the assumption that the components are the same for all the elements of all three modes. We motivated in Section 1 that it can be expected that this assumption does not hold for many empirical data sets. The bread sensory data, which were previously analyzed with CP,[23] constitute a good example. In particular, our reanalysis of these data showed that some of the variables are used in different way by the panelists.

The second aim of this paper was to investigate whether the outlyingness ranking that promises detection heuristics use can be improved by plugging in other similarity measures. The results of our extensive simulation study clearly

indicate that computing Tucker congruence coefficients per component yields the best results when it comes to correctly ranking the variables according to their outlyingness. Indeed, in Section 5.6, we hypothesized that the congruence coefficient–based measures might have an advantage over RMSD-based measures as the latter is more sensitive to small absolute differences between the loadings. Moreover, we expected that the specific subset of loadings that is considered would make a difference. From the simulation study, it became clear that measuring the similarity per component is more effective than measuring it per variable or on all loadings at once. This finding corresponds nicely to the fact that we are looking for components (i.e., columns of loadings) that can be considered equal across the groups.

We see multiple directions for future research. First, one might argue that a limitation of our simulation was that we put the same amount of error on all the groups. It might be more in line with real data to vary this amount across the groups. However, we think that doing this will not change the main outcome of this simulation study, because computing the congruence per component only accounts for proportional similarity, which should not be affected by such error manipulations across the groups. Thus, we would mostly expect the performance difference between the proportional and absolute measures to increase.

Second, the simulation study only contained conditions in which the error level was kept constant across the variables of a group. In empirical data, it may occur that the amount of error varies considerably across variables. In those cases, the added value of other dimension reduction techniques that allow for such error differences could be investigated. For instance, factor analysis[58,59] explicitly models such differences in error variance so that they do not bias the loadings and, thus, cannot affect the outlyingness ranking.

Third, we only considered data sets consisting of two groups. However, in principle, our approach is not limited to this setting. In case the data consist of more than two groups, the number of groups can be reduced by first applying clusterwise SCA,[45] and afterward, if still more than two clusters remain, the heuristics can be applied in a pairwise manner on cluster pairs.[27] This clusterwise SCA approach was illustrated in the application.

Fourth, the outlying variable detection approach is not confined to two-mode data but can also be useful to trace the discrepancies across any two multiway data groups that are variable coupled and are simultaneously analyzed. For example, clusterwise PARAFAC[6] has been proposed to model heterogeneity in the underlying components of three-way three-mode data, and in this context, the outlyingness ranking can probably be applied to further investigate which variables cause that heterogeneity. Linked mode PARAFAC-PCA analysis constitutes another example. This technique simultaneously models a three-way three-mode and two-way two-mode data group that shares one mode, using PARAFAC and PCA respectively, while constraining the component scores of the shared mode to be the same in both models. Again, the outlying variable heuristic can help find which elements of the shared mode behave differently in the two considered groups.

ACKNOWLEDGMENTS

The research leading to the results reported in this paper was supported by the Research Fund of KU Leuven (C14/19/054). For the simulations, we used the infrastructure of the VSC—Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government—department EWI.

ORCID


Sopiko Gvaladze https://orcid.org/0000-0003-3658-4426

REFERENCES

1. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res*. 2014;238(2):391-403.
2. Van Mechelen I, Ceulemans E. Component- and factor-based models for data fusion in the behavioral sciences. *Proc IEEE*. 2015;103(9):1621-1634.
3. Van Deun K, Smilde AK, Thorrez L, Kiers HAL, Van Mechelen I. Identifying common and distinctive processes underlying multiset data. *Chemom Intel Lab Syst*. 2013;129:40-51.
4. Ozcanli F, Ceulemans E, Hermans D, Claes L, Mesquita B. Obsessions across two cultures: a comparison of Belgian and Turkish nonclinical samples. *Front Psychol*. 2019;10:657.
5. Bro R, Qannari EM, Kiers HA, Næs T, Frøst MB. Multi-way models for sensory profiling data. *J Chemometr*. 2008;22(1):36-45.
6. Wilderjans TF, Ceulemans E. Clusterwise PARAFAC to identify heterogeneity in three-way data. *Chemometr Intell Lab Syst*. 2013;129:87-97.
7. De Roover K, Ceulemans E, Timmerman ME. How to perform multiblock component analysis in practice. *Behav Res Methods*. 2012;44(1):41-56.

8. Timmerman ME, Kiers HA. Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*. 2003;68(1):105-121.
9. Kuppens P, Ceulemans E, Timmerman ME, Diener E, Kim-Prieto C. Universal intracultural and intercultural dimensions of the recalled frequency of emotional experience. *J Cross Cult Psychol*. 2006;37(5):491-515.
10. Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol Bull*. 1989;105(2):456-466.
11. Meredith W, Teresi JA. An essay on measurement and factorial invariance. *Med Care*. 2006;44(11):S69-S77.
12. Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*. 1993;58(4):525-543.
13. Widaman KF, Reise SP, Bryant KJ, Windle M. Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In: *The Science of Prevention: Methodological Advances From Alcohol and Substance Abuse Research*. Washington, DC: American Psychological Association; Ariel; 1997:281-324.
14. Holland P, Wainer H. *Differential Item Functioning*. New York: Routledge; 2012.
15. Walker C. What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *J Psychoeduc Assess*. 2011;29(4):364-376.
16. Krysinska K, De Roover K, Bouwens J, et al. Measuring religious attitudes in secularized Western European context: a psychometric analysis of the Post-Critical Belief Scale. *Int J Psychol Relig*. 2014;24(4):263-281.
17. Ceulemans E, Hubert M, Rousseeuw P. Robust multilevel simultaneous component analysis. *Chemom Intel Lab Syst*. 2013;129:33-39.
18. Hahs-Vaughn DL. *Applied Multivariate Statistical Concepts*. New York, NY: Routledge; 2016.
19. Næs T, Brockhoff PB, Tomic O. *Statistics for Sensory and Consumer Science*. United Kingdom: John Wiley & Sons; 2011.
20. Dahl T, Tomic O, Wold JP, Næs T. Some new tools for visualising multi-way sensory data. *Food Qual Prefer*. 2008;19(1):103-113.
21. De Roover K, Timmerman ME, Van Mechelen I, Ceulemans E. On the added value of multiset methods for three-way data analysis. *Chemometr Intell Lab Syst*. 2013;129:98-107.
22. Wilderjans TF, Ceulemans E, Meers K. CHull: a generic convex-hull-based model selection method. *Behav Res Methods*. 2013;45(1):1-15.
23. Bro R. *Multi-Way Analysis in the Food Industry-Models, Algorithms, and Applications*. Denmark: Ph.D. Thesis, University of Amsterdam; 1997.
24. Bro R. PARAFAC. Tutorial and applications. *Chemometr Intell Lab Syst*. 1997;38(2):149-171.
25. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*. 1970;16:1-84.
26. De Roover K, Timmerman ME, Ceulemans E. How to detect which variables are causing differences in component structure among different groups. *Behav Res Methods*. 2017;49(1):216-229.
27. Gvaladze S, De Roover K, Tuerlinckx F, Ceulemans E. Detecting which variables alter component interpretation across multiple groups: a resampling-based method. *Behav Res Methods*. 2019;1-28.
28. Abdi H. Congruence: congruence coefficient, *Rv* coefficient, and mantel coefficient. *Encyclopedia of Research Design*. 2010;222-229.
29. Davenport EC Jr. Significance testing of congruence coefficients: a good idea? *Educ Psychol Meas*. 1990;50(2):289-296.
30. Krzanowski WJ. Between-groups comparison of principal components. *J Am Stat Assoc*. 1979;74(367):703-707.
31. Tucker LR. *A Method for Synthesis of Factor Analysis Studies*. Princeton NJ: (No. PRS-984). Educational Testing Service; 1951.
32. Escoufier Y. Le traitement des variables vectorielles. *Biometrics*. 1973;751-760.
33. Josse J, Pagès J, Husson F. Testing the significance of the *Rv* coefficient. *Comput Stat Data Anal*. 2008;53(1):82-91.
34. Smilde AK, Kiers HA, Bijlsma S, Rubingh CM, Van Erk MJ. Matrix correlations for high-dimensional data: the modified *Rv*-coefficient. *Bioinformatics*. 2008;25(3):401-405.
35. Björck Å, Golub GH. Numerical methods for computing angles between linear subspaces. *Math Comput*. 1973;27(123):579-594.
36. Hitzer E. Angles between subspaces. arXiv preprint arXiv:1306.1629 2013.
37. Caulfield MB, Riggs DS. The assessment of dating aggression: empirical evaluation of the Conflict Tactics Scale. *J Interper Violence*. 1992;7(4):549-558.
38. Levine MS. *Canonical Analysis and Factor Comparison*. Newbury Park, California: SAGE Publications; 1977.
39. Browne MW. On oblique Procrustes rotation. *Psychometrika*. 1967;32(2):125-132.
40. Gower JC. Generalized Procrustes analysis. *Psychometrika*. 1975;40(1):33-51.
41. Lorenzo-Seva U. Promin: a method for oblique factor rotation. *Multivariate Behav Res*. 1999;34(3):347-365.
42. Kiers H, Simplimax A. Oblique rotation to an optimal target with simple structure. *Psychometrika*. 1994;59(4):567-579.
43. Hendrickson AE, White PO. Promax: a quick method for rotation to oblique simple structure. *Br J Stat Psychol*. 1964;17(1):65-70.
44. De Roover K, Vermunt JK, Timmerman ME, Ceulemans E. Mixture simultaneous factor analysis for capturing differences in latent variables between higher level units of multilevel data. *Struct Equ Modeling*. 2017;24(4):506-523.
45. De Roover K, Ceulemans E, Timmerman ME, Vansteelandt K, Stouten J, Onghena P. Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychol Methods*. 2012;17(1):100-119.
46. Yuan S, De Roover K, Dufner M, Denissen JJ, Van Deun K. Revealing subgroups that differ in common and distinctive variation in multi-block data: clusterwise sparse simultaneous component analysis. *Soc Sci Comput Rev*. 2019:432-448, 089443931988844.
47. Van Deun K, Thorrez L, Coccia M, et al. Weighted sparse principal component analysis. *Chemom Intel Lab Syst*. 2019;195:103875.
48. Ceulemans E, Wilderjans TF, Kiers HA, Timmerman ME. Multilevel simultaneous component analysis: a computational shortcut and software package. *Behav Res Methods*. 2016;48(3):1008-1020.
49. Jolliffe IT. *Principal Component Analysis*. New York: Springer-Verlag; 2002.

50. Kiers HA, Ten Berge JM. Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *Brit J Math Stat Psy*. 1994;47(1):109-126.
51. Cattell RB. The scree test for the number of factors. *Multivariate Behav Res*. 1966;1(2):245-276.
52. Ceulemans E, Kiers HA. Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. *Brit J Math Stat Psy*. 2006;59(1):133-150.
53. Horn JL. A rationale and test for the number of factors in factor analysis. *Psychometrika*. 30(2):179-185.
54. Chan W, Ho RM, Leung K, Chan DKS, Yung YF. An alternative method for evaluating congruence coefficients with Procrustes rotation: a bootstrap procedure. *Psychol Methods*. 1999;4(4):378-402.
55. Lorenzo-Seva U, Ten Berge JM. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*. 2006;2(2):57-64.
56. Kaiser HF. The VARIMAX criterion for analytic rotation in factor analysis. *Psychometrika*. 1958;23(3):187-200.
57. Wedin PÅ. On angles between subspaces of a finite dimensional inner product space. In: *Matrix Pencils*. Berlin, Heidelberg; 1983: 263-285.
58. Lawley DN, Maxwell AE. Factor analysis as a statistical method. *J R Stat Soc, Ser D Stat*. 1962;12(3):209-229.
59. Comrey AL, Lee HB. *A First Course in Factor Analysis*. Hillsdale: Erlbaum: Psychology Press; 2013.

How to cite this article: Gvaladze S, De Roover K, Tuerlinckx F, Ceulemans E. Detecting outlying variables in multigroup data: A comparison of different loading similarity coefficients. *Journal of Chemometrics*. 2021;35: e3233. https://doi.org/10.1002/cem.3233