Behav Res (2014) 46:576–587
DOI 10.3758/s13428-013-0374-6

Performing DISCO-SCA to search for distinctive and common information in linked data

Martijn Schouteden · Katrijn Van Deun · Tom F. Wilderjans · Iven Van Mechelen

Published online: 1 November 2013
© Psychonomic Society, Inc. 2013

Abstract Behavioral researchers often obtain information about the same set of entities from different sources. A main challenge in the analysis of such data is to reveal, on the one hand, the mechanisms underlying all of the data blocks under study and, on the other hand, the mechanisms underlying a single data block or a few such blocks only (i.e., common and distinctive mechanisms, respectively). A method called DISCO-SCA has been proposed by which such mechanisms can be found. The goal of this article is to make the DISCO-SCA method more accessible, in particular for applied researchers. To this end, first we will illustrate the different steps in a DISCO-SCA analysis, with data stemming from the domain of psychiatric diagnosis. Second, we will present in this article the DISCO-SCA graphical user interface (GUI). The main benefits of the DISCO-SCA GUI are that it is easy to use, strongly facilitates the choice of model selection parameters (such as the number of mechanisms and their status as being common or distinctive), and is freely available.

Keywords Common and distinctive · Simultaneous component analysis · Rotation · Linked data · Graphical user interface

In behavioral research, it often occurs that information about the same set of entities (e.g., items, persons, situations, . . .) is obtained from many different sources (e.g., different populations, moments in time, psychological tests, . . .). In the field of personality psychology, for instance, Rossier, de Stadelhofen, and Berthoud (2004) subjected the same group of subjects to two different personality questionnaires in order to compare the aspects of personality measured by the two questionnaires. In this way, Rossier et al. obtained two Person × Item data blocks, one per questionnaire, with information about the same persons. Another example can be found in the field of behavioral genetics, where Nishimura et al. (2007) investigated the expression profile of a set of genes in three different populations, these being (a) males with autism spectrum disorder (ASD) due to a fragile X mutation, (b) males with ASD due to a 15q11-q13 duplication, and (c) nonautistic controls. In this way, a data set was obtained consisting of three Person × Gene data blocks with information about the same set of genes. Depending on which set of entities are common to the different data blocks, such data sets will be referred to as *row-/object-wise* (e.g., first example) or *column-/variable-wise* (e.g., second example) linked data (see Fig. 1 for a graphical representation).

Major challenges in the analysis of such linked data are (1) revealing the underlying behavioral mechanisms in the whole data set and (2) disentangling therein the mechanisms underlying all of the data blocks under study (i.e., common mechanisms) and the mechanisms underlying a single or a few data blocks only (i.e., distinctive mechanisms). For instance, Rossier et al. (2004) wanted to discover which aspects of personality were measured by both questionnaires, as well as which aspects were specific for each questionnaire; Nishimura et al. (2007) were interested in biological functions that were characteristic for both autism groups but not for the control group. Note that we use the term *mechanism* generically to denote the underlying causes of variation in the data. Depending on the application, these causes may be vague, as is often the case with such latent variables as personality traits, and the results of the data analysis may be considered summaries of the variables (Fabrigar, Wegener, MacCallum, & Strahan, 1999). However, in other cases the

M. Schouteden · K. Van Deun (✉) · T. F. Wilderjans · I. Van Mechelen
Research Group Quantitative Psychology and Individual Differences, KU Leuven, Tiensestraat 102, bus 3713, 3000 Leuven, Belgium
e-mail: katrijn.vandeun@ppw.kuleuven.be

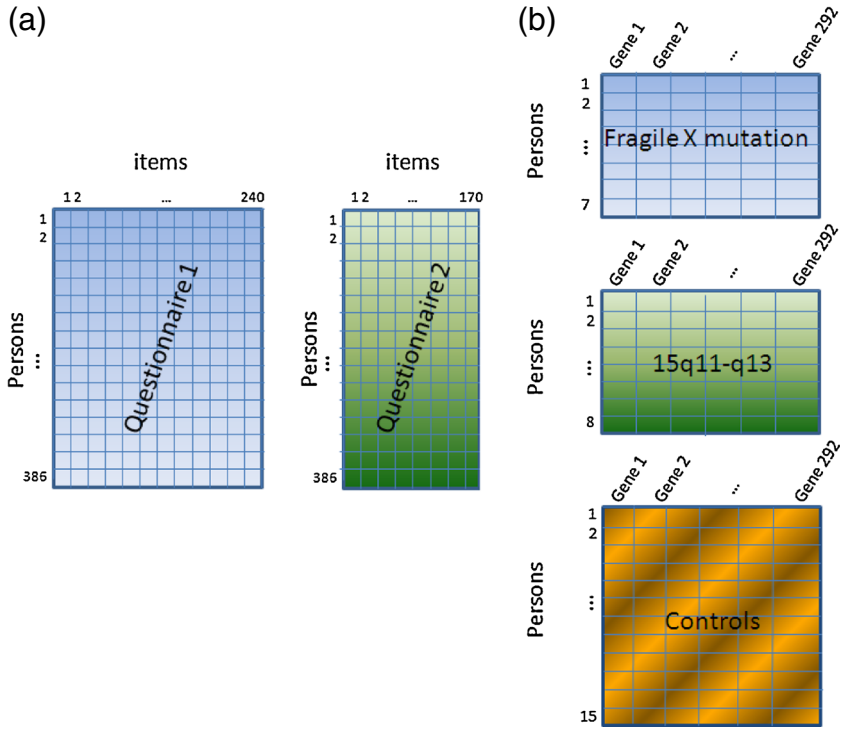

Fig. 1 Graphical representation of linked data consisting of different data blocks with information about the same set of persons (row-/object-wise linked data; panel **a**), or, the same set of genes (column-/variable-wise linked data; panel **b**)

underlying causes of variation have a tangible physical or chemical nature that is directly reflected by the results of the data analysis (Tauler, Smilde, & Kowalski, 1995).

Schouteden, Van Deun, Pattyn, and Van Mechelen (2013) have recently proposed a method, called *DISCO-SCA*, by which such distinctive and common mechanisms can be revealed (see also Van Deun et al., 2012). To facilitate the use of DISCO-SCA in empirical practice, we developed freely available software, both as a standalone version for Microsoft Windows and as a MATLAB version. These can be downloaded from http://ppw.kuleuven.be/okp/software/disco-sca/.

The remainder of this article is organized in four sections. In section "DISCO-SCA", the DISCO-SCA model, algorithm, model selection, and related methods are briefly discussed. In section "The DISCO-SCA process", the full DISCO-SCA process is outlined and illustrated on a public available data set stemming from the field of psychiatric diagnosis. In section "The DISCO-SCA program", the DISCO-SCA program is discussed. In section "Conclusion", we present a conclusion.

DISCO-SCA

To find the common and distinctive mechanisms underlying linked data, DISCO-SCA operates in two steps. In the first step, the linked data are analyzed by means of simultaneous component (SCA) methods (see, e.g., Kiers & ten Berge, 1989; ten Berge, Kiers, & Van der Stel, 1992; for a review, see Van Deun, Smilde, van der Werf, Kiers, & Van Mechelen, 2009). SCA is a family of component methods that have in particular been developed for the analysis of linked data. These methods typically reveal a small number of simultaneous components that maximally account for the variation in the whole data set. However, the components in question typically reflect a mix of common and distinctive information. Therefore, in a second step, DISCO-SCA disentangles the two kinds of information by making use of the rotational freedom of the simultaneous components. In the following two subsections, both steps will be discussed in more detail. Next, we will devote a subsection to the algorithm to estimate the different DISCO-SCA parameters. We will conclude this section with a subsection discussing the problem of model selection. To ease the explanation, without loss of generality, we will focus on a model for column-wise linked data blocks.

Step 1: Simultaneous component analysis

Given Q components, SCA decomposes K linked $I_k \times J$ data blocks $\mathbf{X}_k$ ($k=1, \ldots, K$), as follows:

$$\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix} = \begin{bmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_K \end{bmatrix} \mathbf{P}' + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_K \end{bmatrix}, \quad (1)$$

with $\mathbf{T}_k$ being a block-specific $I_k \times Q$ score matrix, $\mathbf{P}$ a common $J \times Q$ loading matrix, and $\mathbf{E}_k$ a block-specific $I_k \times J$ matrix of residuals. Note that the use of a common $\mathbf{P}$ shows that the data blocks are linked column-wise. Eq. 1 is equivalent to

$$\mathbf{X}_{\text{conc}} = \mathbf{T}_{\text{conc}}\mathbf{P}' + \mathbf{E}_{\text{conc}}, \tag{2}$$

with $\mathbf{X}_{\text{conc}}$ denoting the $(\sum_{k=1}^{K} I_k \times J)$ matrix that is obtained by concatenating all data blocks $\mathbf{X}_k$, and $\mathbf{T}_{\text{conc}}$ the $(\sum_{k=1}^{K} I_k \times Q)$ matrix of component scores resulting from concatenating all $\mathbf{T}_k$, and $\mathbf{E}_{\text{conc}}$ the $(\sum_{k=1}^{K} I_k \times J)$ matrix of residuals resulting from the concatenation of all $\mathbf{E}_k$. To identify the model, the component scores are constrained, for example such that they are orthonormal: $\mathbf{T}'_{\text{conc}}\mathbf{T}_{\text{conc}} = \mathbf{I}$.

The scores and loadings can be estimated by minimizing the following objective function:

$$\min_{\mathbf{T}_{\text{conc}},\mathbf{P}} \|\mathbf{X}_{\text{conc}} - \mathbf{T}_{\text{conc}}\mathbf{P}'\|^2 \text{ such that } \mathbf{T}'_{\text{conc}}\mathbf{T}_{\text{conc}} = \mathbf{I}, \tag{3}$$

with the notation $\|\mathbf{Z}\|^2$ indicating the sum of squared elements of the matrix $\mathbf{Z}$.[1] There is no unique solution to Eq. 3: Let $\mathbf{T}_{\text{conc}}$ and $\mathbf{P}$ be a solution; then, the (orthogonal) rotation $\mathbf{T}^*_{\text{conc}} = \mathbf{T}_{\text{conc}}\mathbf{B}$ and $\mathbf{P}^* = \mathbf{P}\mathbf{B}$, with $\mathbf{B}'\mathbf{B} = \mathbf{I} = \mathbf{B}\mathbf{B}'$ orthonormal is a solution of Eq. 3, too. Note that this solution corresponds to the SCA-P approach to simultaneous component analysis (Kiers & ten Berge, 1989; Timmerman & Kiers, 2003).

The results of the simultaneous component analysis strongly depend on how the data were preprocessed. Often, the different data blocks are corrected for differences in the offset and scale of the variables, to give equal weight to each variable. However, in the case of variable-wise linked data, it may be needed to center and/or scale the variables over all data blocks simultaneously: Centering per block removes differences in the means between blocks, and scaling per block removes differences in intrablock variability that may exist between blocks. If such differences are artificial or of no interest, it may indeed be advised to remove them prior to the simultaneous component analysis by centering and/or scaling per block; otherwise, if they are meaningful and of interest, the variables should be centered and/or scaled over the blocks. See Timmerman and Kiers (2003) and Bro and Smilde (2003) for a more elaborate discussion of centering and scaling. Furthermore, in the case that the data blocks differ considerably in size, the results may be dominated by the largest data block (van den Berg et al., 2009; Wilderjans, Ceulemans, Van Mechelen, & van den Berg, 2009). A possible strategy then could be to scale each data block to the sum of squares 1 (see, e.g., Timmerman, 2006; Van Deun et al., 2009; and Wilderjans, Ceulemans, & Van Mechelen, 2009, for more information about preprocessing and weighting linked data blocks).

[1] In the implementation, the scores are rescaled to $\mathbf{T}_{\text{conc}} = (\sum_k I_k)^{(-1/2)}\mathbf{T}_{\text{conc}}$ and the loadings to $\mathbf{P} = (\sum_k I_k)^{1/2}\mathbf{P}$; then, for standardized variables, the loadings coincide with the correlations between the variables and component scores (Van Deun, Wilderjans, van den Berg, Antoniadis, & Van Mechelen, 2011).

Step 2: Rotation

A mechanism that is distinctive for a single data block $\mathbf{X}_k$ is defined as a simultaneous component with block-specific scores equal to zero for all data blocks except data block $\mathbf{X}_k$; a mechanism that is distinctive for more than one data block is defined as a simultaneous component with block-specific scores equal to zero for all data blocks except those for which it is distinctive. For example, in the case of three data blocks, a component that is distinctive for the first two data blocks is defined as a component with zero scores for the third data block. The scores of the common components do not have such prespecified zero parts (Schouteden et al., 2013). Note that in the case of data that are linked row-wise (i.e., the objects are the shared mode between the blocks), distinctive components are defined by zero *loadings* for the blocks that the component does not underlie. In general, the solution obtained by minimizing Eq. 3 does not contain a clear common/distinctive structure, resulting in components capturing a mix of common and distinctive information. DISCO-SCA uses the rotational freedom of SCA to address this problem, and (orthogonally) rotates the scores of the simultaneous components toward a clear common/distinctive structure (denoted by $\mathbf{T}^{\text{target}}_{\text{conc}}$). An example of such a target structure, for a case with $K=2$ data blocks and $Q=3$ components, with the first component being distinctive for the first data block, the second distinctive for the second data block, and the third common, is:

$$\mathbf{T}^{target}_{conc} = \begin{bmatrix} \mathbf{T}^{target}_1 \\ - \; - \; - \\ \mathbf{T}^{target}_2 \end{bmatrix} = \begin{bmatrix} * & 0 & * \\ \vdots & \vdots & \vdots \\ * & 0 & * \\ - & - & - \\ 0 & * & * \\ \vdots & \vdots & \vdots \\ 0 & * & * \end{bmatrix}, \tag{4}$$

where * denotes an unspecified entry.

To find the rotation matrix $\mathbf{B}$ that rotates the component scores $\mathbf{T}_{\text{conc}}$ toward the target, the following least-squares optimization criterion is introduced:

$$\min_{\mathbf{B}} \left\| \mathbf{W} \circ \left(\mathbf{T}_{\text{conc}}\mathbf{B} - \mathbf{T}^{\text{target}}_{\text{conc}} \right) \right\|^2, \tag{5}$$

subject to $\mathbf{B}'\mathbf{B} = \mathbf{I} = \mathbf{B}\mathbf{B}'$; matrix $\mathbf{W}$ denotes a weight matrix with ones in the positions corresponding to the zeroes in $\mathbf{T}^{\text{target}}_{\text{conc}}$, and with zeroes elsewhere (Browne, 1972); $\circ$ denotes the element-wise or Hadamard product. The rotated component loadings, which are the same for all data blocks, then equal $\mathbf{PB}$; the rotated scores equal $\mathbf{T}_{\text{conc}}\mathbf{B}$.

The target rotation criterion (Eq. 5) has no unique solution when two or more components have the same status (e.g., two distinctive components for the first data block and/or two common components): such components result in identical columns in the target and weight matrices, and it can be shown that any orthogonal rotation of these components yields the same value for the target rotation criterion in Eq. 5; hence, there is no unique optimal solution, but there are many different ones. DISCO-SCA deals with this identification problem by first finding the overall rotation matrix by solving Eq. 5, and subsequently subjecting the loadings of each set of components with the same status to a VARIMAX (Kaiser, 1958) or EQUAMAX (Saunders, 1962) rotation. The rotation of the loadings is compensated for by subsequently counterrotating the component scores. The choice of VARIMAX or EQUAMAX is made in view of getting closer to a simple structure for the (subset of) loadings under study that may facilitate the interpretation of the components.

Algorithm

A solution to the objective function in Eq. 3 is the singular value decomposition of $\mathbf{X}_{\text{conc}}$:

$$\mathbf{X}_{\text{conc}} = \mathbf{USV}', \tag{6}$$

with $\mathbf{U}$ and $\mathbf{V}$ being orthonormal ($\mathbf{U}'\mathbf{U}=\mathbf{I}=\mathbf{V}'\mathbf{V}$) and $\mathbf{S}$ being a diagonal matrix containing the singular values ranked from largest to smallest. For a solution with Q simultaneous components, the component score matrix $\mathbf{T}_{\text{conc}}$ and the loading matrix $\mathbf{P}$ equal

$$\begin{aligned} \mathbf{T}_{\text{conc}} &= \mathbf{U}_Q \\ \mathbf{P} &= \mathbf{V}_Q\mathbf{S}_Q, \end{aligned} \tag{7}$$

with $\mathbf{U}_Q$ and $\mathbf{V}_Q$ denoting the first Q singular vectors, and $\mathbf{S}_Q$ being a diagonal matrix containing the first Q singular values. The minimization of the rotation criterion (Eq. 5) is presented in the Appendix.

Model selection

When applying DISCO-SCA to given data, two model selection problems need to be sorted out: The first problem pertains to selecting the number of simultaneous components Q that underlie the data, and the second to determining the status of the components (i.e., finding the optimal target matrix).

Regarding the first model selection problem, Van Deun et al. (2009) proposed selecting the simultaneous components that explain a sizeable amount of variance in *at least one* data block. Here, we define *sizeable* for a component in a block as being more than the 95th percentile of the empirical distribution of variance accounted for (VAF), for that particular combination of block and component, resulting from a resampling strategy; note that this model selection procedure is a variant of parallel analysis (Buja & Eyuboglu, 1992; Horn, 1965; Peres-Neto, Jackson, & Somers, 2005). The stability of the thus selected number of components can be assessed by means of a bootstrap analysis; note that stability selection has shown promising results in related problems (Meinshausen & Buhlmann, 2010). For an example of this method, we refer the reader to section "Selecting the number of simultaneous components", and for more details, to Schouteden et al. (2013).

Regarding the second model selection problem, given a fixed number of components, many possible target structures can be constructed. Schouteden et al. (2013) proposed an exhaustive procedure that consists of measuring for each target the deviation of the observed solution from the ideal and selecting the target with the lowest deviation. The ideal for a target is defined by defining each of its components as follows: An ideal distinctive component for one or more data block(s) is defined as a component with a sum of squared component scores of 0 (i.e., all component scores are 0) in the remaining block(s); an ideal common component is defined to have equal sums of squared component scores within each block; this sum of squared block-specific component scores is equal to $c_q = K^{-1}\sum_k\sum_{i_k} t_{i_kq}^2$ for all $k=1, \ldots, K$. An illustration for a target pertaining to two data blocks and three components in which the first component is defined to be specific for the first data block, the second to be specific for the second data block, and the third to be common to both data blocks is shown in Table 1. The ideal is represented in the two last rows of the table, with the ideal for the common component being derived from the component scores that are observed after rotation to the target. The overall deviation is computed by summing the deviations from 0 or c_q over all components and all data blocks, with the deviation from 0 being measured by $\sum_k \sum_{i_k}\left(0-t_{i_kq}^2\right)^2$ and the deviation from c_q by $\sum_k\left(c_q-\sum_{i_k} t_{i_kq}^2\right)^2$. The stability of this model selection heuristic can be assessed in a bootstrap analysis (for more details, see Schouteden et al., 2013).

It should be noted that the present model selection procedure may yield too few components, due to selecting the number of components on the basis of the results of the (unrotated) simultaneous component analysis: The unrotated components sequentially account for an optimal amount of variation in the concatenated data, whereby distinctive components may be missed. This may be solved for if the *rotated* components that account for a sizeable amount of variance in at least one block are retained. However, the problem is that for variable-wise coupled data, both the block-specific component scores and loadings are no longer orthogonal after

Table 1 Calculating the deviation of the observed solution to the ideal for a target consisting of one specific component for the first data block $\mathbf{X}_1$, one specific component for the second data block $\mathbf{X}_2$, and one common component

		$\mathbf{X}_1$ Specific	$\mathbf{X}_2$ Specific	Common
Block 1	1	t_{11}^2	t_{12}^2	t_{13}^2
	2	t_{21}^2	t_{22}^2	t_{23}^2
	…	…	…	…
	I_1	$t_{I_1 1}^2$	$t_{I_1 2}^2$	$t_{I_1 3}^2$
	Sum 1	c_{11}	c_{12}	c_{13}
Block 2	1	t_{12}^2	t_{12}^2	t_{13}^2
	2	t_{22}^2	t_{22}^2	t_{23}^2
	…	…	…	…
	I_2	$t_{I_2 2}^2$	$t_{I_2 2}^2$	$t_{I_2 3}^2$
	Sum 2	c_{21}	c_{22}	c_{23}
	Sums 1+2	$Kc_1=c_{11}+c_{21}$	$Kc_2=c_{12}+c_{22}$	$Kc_3=c_{13}+c_{23}$
	Ideal $\mathbf{X}_1$	/	0	c_3
	Ideal $\mathbf{X}_2$	0	/	c_3

The upper parts of the table contain the observed squared component scores and their sum, as obtained for each data block after rotation to the target. The two last rows in the bottom correspond to the ideal for the target; a forward slash "/" indicates that no ideal applies to that particular combination of data block and component.

rotation. Therefore, the VAF by a component in a block is no longer a pure contribution of that component to the overall VAF in the block but also depends on other components.

Related methods

Besides DISCO-SCA, some other methods have been proposed that deal with the problem of finding common and specific components in multiset data. For the limiting case of two data blocks, the generalized singular value decomposition (GSVD) has been suggested as a method for finding common and specific components (Alter, Brown, & Botstein, 2003); Van Deun et al. (2012) showed how to properly use the GSVD such that it becomes a data-approximation method that also performs well when retaining only a few components. This adapted GSVD returns a rotation of the DISCO-SCA solution in the case of variable-wise linked data and is equivalent to SCA-IND (Timmerman & Kiers, 2003). For object-wise coupled data with two or more data blocks, OnPLS (Lofstedt & Trygg, 2011) has been proposed. This method yields a set of components for each data block with specific components that are uncorrelated to the common components. This is different from simultaneous component analysis approaches that yield a single set of components shared between all data blocks: Common components are clearly the same components for all data blocks, whereas the distinctive components are clearly absent in the data block(s) that they do not underlie. For the case of variable-wise linked data with preferably many blocks, a cluster-wise simultaneous component analysis (De Roover et al., 2012) with common and cluster-specific simultaneous components (De Roover, Timmerman, Mesquita, & Ceulemans, 2013) has been proposed. The latter model is inspired by the basic idea of DISCO-SCA to have zero scores in the specific components for the parts that should *not* underlie a group of data blocks. Unlike DISCO-SCA, the method first clusters the data blocks into a few groups and imposes the scores to be exactly equal to 0. Also, the common and specific components are uncorrelated at the level of the individual blocks.

The DISCO-SCA process

In this section, we will discuss the main steps in DISCO-SCA, which are (1) preprocessing of the data, (2) choosing the optimal number of simultaneous components, and (3) defining the status of the components (this is selecting the optimal target matrix). We will illustrate with a publicly available data set stemming from a study in the field of psychiatry (Mezzich & Solomon, 1980). In this study, 22 experienced psychiatrists were asked to rate how well certain

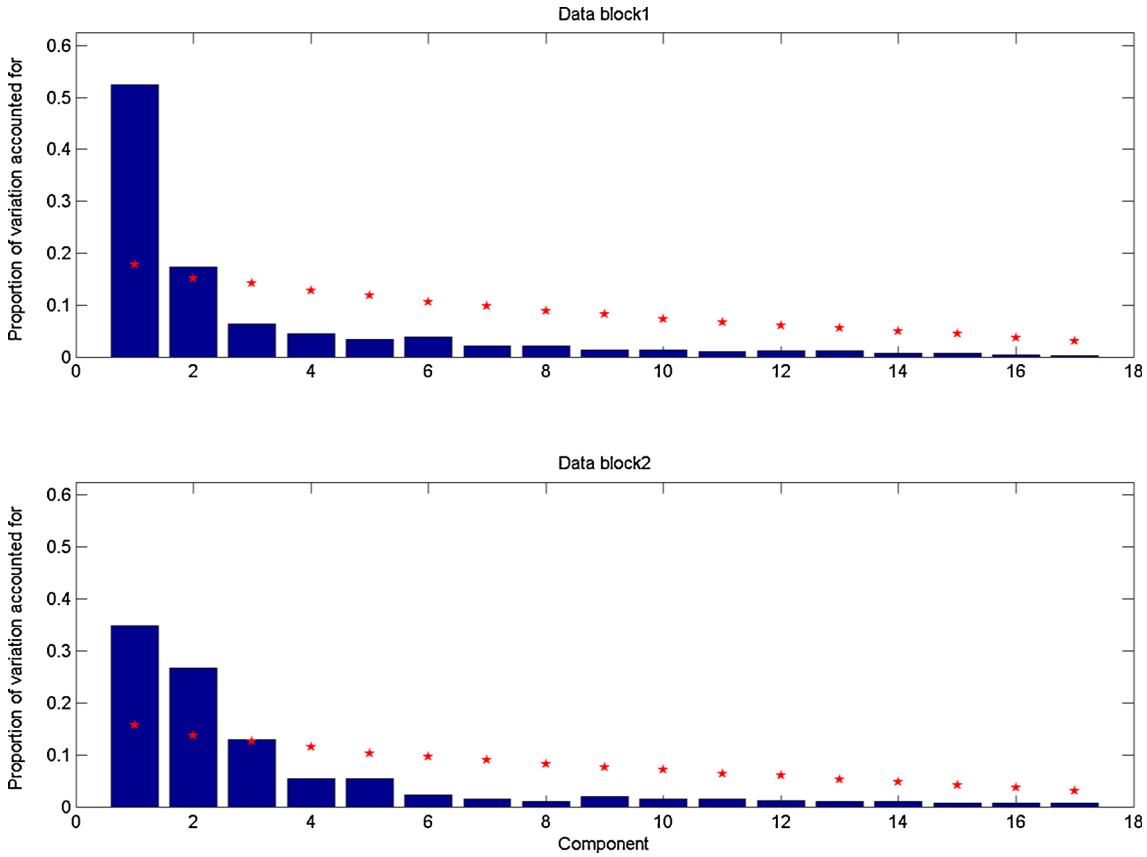

Fig. 2 Proportions of variance accounted for by each simultaneous component in each block of the psychiatric data (upper panel=manic-depressed patients; lower panel=schizophrenic patients). The stars indicate the critical noise values obtained with parallel analysis

symptoms matched certain archetypal psychiatric patients on a 7-point scale ranging from 0 (*not at all*) to 6 (*very applicable*). The study included four archetypal psychiatric patients—that is, paranoid schizophrenics, simple schizophrenics, manic manic-depressed, and depressed manic-depressed patients—and 17 symptoms, including "anxiety," "hostility," "guilt feelings," and so forth (see Table 3 below). This resulted in four Psychiatrist×Symptom data blocks, one for each archetypal patient. For illustrative reasons, we grouped together the data blocks for the manic and depressed

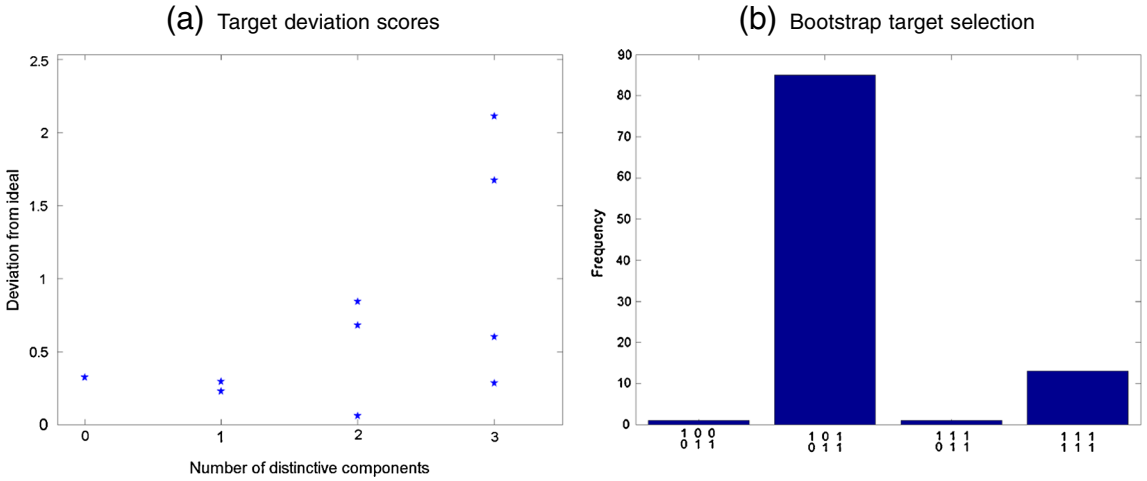

Fig. 3 (**a**) Target deviation scores: Deviation of the observed from the ideal sums of squared component scores as a function of the number of distinctive components for the ten possible targets for three components. (**b**) Bootstrap of the target selection: For each target matrix, the number of bootstrap samples (out of 100 bootstrap replications) resulting in selection of the target is plotted. The target is labeled by a binary Block×Component matrix indicating whether the component is present (score 1) or absent (score 0) in the block

Table 2 Block-specific sums of squared component scores after rotation (before-rotation scores appear between parentheses)

	$C_{\text{depressed}}$	$C_{\text{schizophrenic}}$	C_{common}
$X_{\text{depressed}}$	.95 (.65)	**.13** (.45)	.40 (.38)
$X_{\text{schizophrenic}}$	**.05** (.35)	.87 (.55)	.60 (.62)

Sums of squared component scores that ideally should be zero are put in bold. $C_{\text{depressed}}$ and $C_{\text{schizophrenic}}$ denote the distinctive components for the manic-depressed and the schizophrenic patients, respectively; C_{common} denotes the common component.

manic-depressed patients, on the one hand, and those for the paranoid and simple schizophrenics on the other hand. This resulted in two Psychiatrist × Symptom data blocks, one for the manic-depressed patients and one for the schizophrenic patients. We will treat these data as being linked variable-wise—that is, by the symptoms. DISCO-SCA will be used to extract the common and distinctive information between these two psychiatric groups. The data set can be found in the folder "Data," located in the directory to which the DISCO-SCA_MATLAB.zip file has been extracted.

Data preprocessing

The DISCO-SCA program offers the following preprocessing procedures and combinations thereof: centering and scaling to sum of squares 1 of the variables per/over all data blocks, and weighting the data blocks to equal sums of squares. The primary aim of the DISCO-SCA analysis of the psychiatric diagnosis data was to reveal common and distinctive sources of variation, rather than to account for between-block differences in the means; therefore, we centered the symptoms per data block. Furthermore, to give all symptoms equal weights in the analysis, and to preserve possibly interesting differences in variability between the data blocks, we chose to scale (to 1) the symptoms jointly across the data blocks.

Table 3 Loadings on the three simultaneous components after rotation

	$C_{\text{depressed}}$	$C_{\text{schizophrenic}}$	C_{common}
Somatic concern	**–.81**	–.28	.23
Anxiety	**–.59**	**–.61**	.14
Emotional withdrawal	**–.77**	**.39**	.21
Conceptual disorganization	.32	–.10	**.78**
Guilt feelings	**–.89**	.03	.15
Tension	**.43**	**–.55**	.33
Mannerisms and posturing	.15	.12	**.70**
Grandiosity	**.75**	**–.60**	.05
Depressive mood	**–.91**	.07	.19
Hostility	**.54**	**–.69**	–.01
Suspiciousness	.18	**–.85**	.22
Hallucinatory behavior	.05	**–.64**	**.49**
Motor retardation	**–.85**	**.35**	.20
Uncooperativeness	**.36**	**–.68**	.24
Unusual thought content	–.03	**–.52**	**.66**
Blunted effect	–.29	**.78**	.24
Excitement	**.80**	**–.45**	.08

$C_{\text{depressed}}$ and $C_{\text{schizophrenic}}$ denote the distinctive components for the manic-depressed and the schizophrenic patients, respectively; C_{common} denotes the common component. Loadings with absolute value ≥.35 have been put in bold.

Selecting the number of simultaneous components

Figure 2 displays for each data block separately the proportion of variance accounted for by each simultaneous component (upper panel, manic-depressed patients; lower panel, schizophrenic patients), along with their critical noise values (for the 95th percentile) obtained from a parallel analysis with 100 samples (see section "Model selection"). From this figure, it appears that only the first three simultaneous components exceed the critical noise level in at least one data block; this suggests that a three-component solution should be retained. In a bootstrap analysis with 50 bootstrap replications, the same number of components was selected in the majority of the cases (i.e., 70 %).

Selecting an optimal target matrix

Given two data blocks and three simultaneous components, ten different target matrices are possible. In Fig. 3, for each of the possible target matrices, the deviation of the observed from the ideal sum of squared component scores (the so-called *deviation score*; see section "Model selection") is plotted as a function of the number of distinctive components. The lowest deviation is obtained for the solution with one distinctive component for the manic-depressed patients, one distinctive component for the schizophrenic patients, and one common component. This solution was selected in 80 % of the cases in a bootstrap analysis with 100 replications, and can therefore be considered to be fairly stable; see Fig. 3.

Interpretation of the results

In Table 2, the block-specific sums of squared component scores before and after rotation are reported for each component and each data block (with the sum of squared component scores that ideally should be zero put in bold). After rotation, the first component is the distinctive component for the depression data block, the second component is distinctive for the schizophrenia data block, and the third component is the common one. From this table, it clearly appears that the DISCO-SCA rotation resulted in components with a

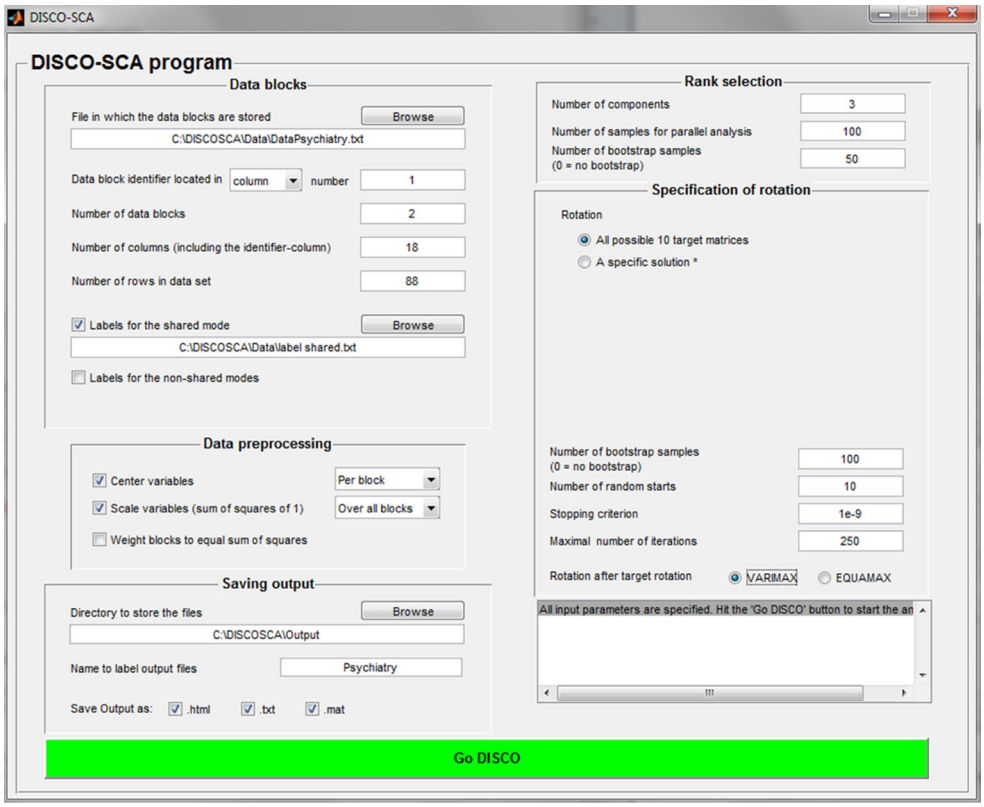

Fig. 4 Screenshot of the DISCO-SCA program, applied to the psychiatry data set described and discussed in section "The DISCO-SCA program"

clearer common/distinctive structure than was present before rotation.[2]

Components can be interpreted on the basis of the highest (in an absolute sense) loadings of the variables (see Table 3). From this table, it clearly appears that the distinctive component for the manic-depressed patients is a bipolar one, which can be labeled as "manic" versus "depressed." The distinctive component for the schizophrenic patients is also bipolar and comprises symptoms of simple versus paranoid schizophrenia. Finally, the common component seems to reflect a general active disturbance of perception/cognition/motor behavior.

The DISCO-SCA program

The MATLAB (in DISCO-SCA_MATLAB.zip) version and the standalone (32- or 64-bit) version for Microsoft Windows of the DISCO-SCA program can be downloaded from http://ppw.kuleuven.be/okp/software/disco-sca/. After setting the current MATLAB directory to the folder that is extracted from DISCO-SCA_MATLAB.zip, the MATLAB version can be launched by typing "DISCO_SCA" in the MATLAB command window, followed by pressing <ENTER>. The standalone version can be launched by double clicking on the DISCO-SCA icon (i.e., DISCO-SCA.exe).

After launching the DISCO-SCA program, the graphical user interface (GUI; displayed in Fig. 4) appears. The GUI consists of one (initially red-colored) push-button called *(NO) GO DISCO*, one message box, and the following five panels: *Data blocks*, *Data preprocessing*, *Rank selection*, *Specification of rotation*, and *Saving output*. Each panel consists of different fields that need to be filled out correctly. In the next five subsections, we will explain for each panel how to do so. We will then discuss how to start the analysis and how to deal with errors.

Data blocks

The file that contains the row- or column-wise linked data, the number of data blocks, the size properties, and, optionally, the file that has the labels for the shared and/or nonshared mode have to be specified in the *Data blocks* panel.

To select the data file, click the appropriate *Browse* button. The data file should be an ASCII file (i.e., a .txt file) containing all data blocks concatenated according to the common mode (i.e., vertically when the variables/columns are common, and horizontally when the objects/rows are common). Each data element should be an integer or real

[2] A *distinctive component* is defined as a component with component scores ideally equal to zero in one or more data blocks and, as a consequence, a sum of squared component scores equal to zero for these data block(s); a common component does not have such prespecified zero parts (see section DISCO-SCA: Step 2: Rotation).

number, with a period as decimal separator. Note that the DISCO-SCA program cannot deal with missing values.

To identify the different data blocks, an extra column (in the case of column-wise linked data blocks) or row (in the case of row-wise linked data blocks) should be added to the data file. Each element of this extra column (row), which is called the *Data block identifier*, is an integer that indicates to which data block the corresponding object (variable) belongs. As an example, a part of the psychiatry data set, in which the data blocks are linked column-wise (i.e., the variables/symptoms are common) is displayed in Fig. 5, with the *Data block identifier* (i.e., extra column) being the first column.

After specifying in which column (row) the *Data block identifier* is located, the user further needs to specify the number of data blocks, the number of columns, and the number of rows of the concatenated data set. Finally, the user is given the opportunity to provide label files for the shared and/or nonshared mode(s) of the data set by clicking on the appropriate *Browse* buttons. Each label file should be an ASCII file containing one column with the labels for the mode in question. When no labels are provided, the DISCO-SCA program will create standard labels (i.e., "obj. 1," "obj. 2," . . ., for the rows, and "var. 1," "var. 2," . . ., for the columns).

Data preprocessing

As we mentioned in section "DISCO-SCA", the DISCO-SCA program provides several options to preprocess the data. These options can be activated by clicking on the appropriate check box. In the case of centering and scaling the variables, the user can choose to do this *per block* or *over all blocks*. In the case that none of the three boxes is checked, the data will not be preprocessed.

Rank selection

In the *Rank selection* panel, the user must first specify the total number of common and distinctive components. Then, the user can set the number of samples for the parallel analysis (default 100, with 0 meaning that no parallel analysis will be performed) and for the bootstrap replications (default 0, implying that no bootstrap analysis is performed). When no parallel analysis is required by the user, a bootstrap analysis for the number of components cannot be performed, because this depends on the critical noise values, as determined by a parallel analysis.

Specification of rotation

After specifying the total number of common and distinctive components in the *Rank selection* panel, the user has to choose—by clicking on the appropriate button—whether he or she wants to rotate the simultaneous components toward *All possible target matrices* or toward *A specific target matrix*. If the latter option is chosen, a table appears in which the rows pertain to the different data blocks and the columns to the different components. Specifying a distinctive component for (a) particular data block(s) can be done by clicking on the cell(s) located in the intersection of the row(s) of the other data block(s) and the column pertaining to the component in question (i.e., a selected cell implies an ideal sum of squared component scores of 0 for the

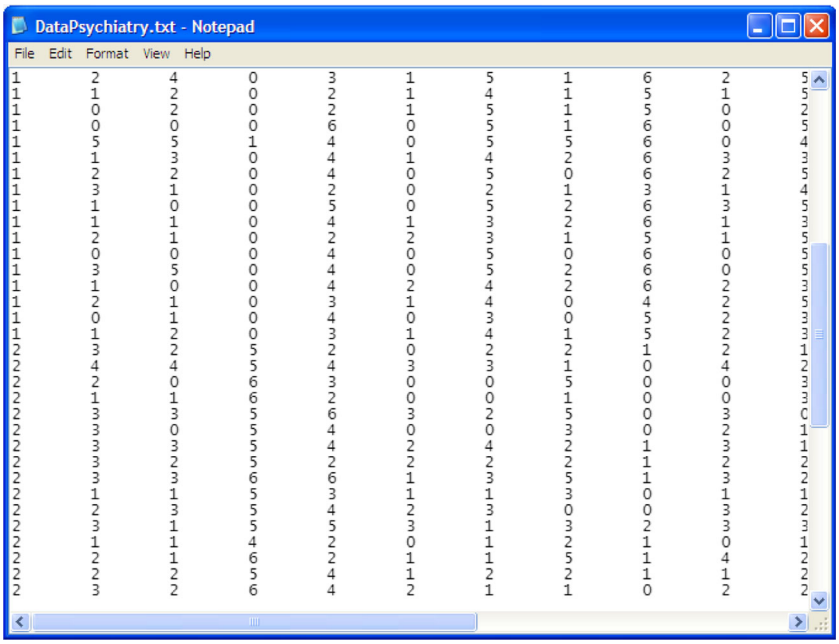

Fig. 5 Screenshot of a part of the psychiatry data set; the first column is the *Data block identifier*. The rows that have the same number in their first column belong to the same data block (the manic-depressed patients are labeled by "1," the schizophrenic patients by "2")

associated data block and component). The Ctrl button on the keyboard can be used to select more cells or to undo the selection of a cell. Cell selection can also be undone by clicking a nonselected cell. Note that when no cell is selected, all components are considered common. Note further that it is not possible to select all cells of one column, as this would imply that the corresponding component does not underlie any data block. An example is given in Fig. 6, where we have chosen to rotate the scores of three simultaneous components toward the target matrix (Eq. 4) with the first and second components being distinctive for the second and first data blocks, respectively, and the third component being common (see section "Step 2: Rotation"). When *All possible target matrices* is chosen, the user has the option to perform a bootstrap analysis, of which the number of bootstrap replications should be specified (the default 0 implying that no bootstrap analysis will be performed).

Finally, the user has to specify the number of random starts (default 10), the stopping criterion (default 10^{-9}), the maximal number of iterations (default 250), and the type of rotation (VARIMAX or default EQUAMAX) for postprocessing the DISCO-SCA output in order to solve the identification problem that occurs when two or more components are of the same type (see section "Step 2: Rotation").

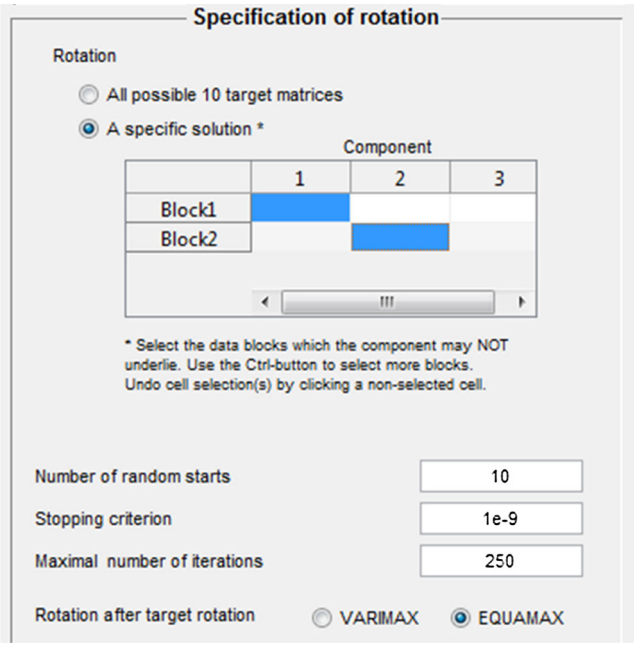

Fig. 6 Screenshot of the *Specification of rotation* panel of the DISCO-SCA program, in which we have chosen to rotate the components toward a target matrix in which the first and second components are distinctive for the second and first data blocks, respectively, and the third component is common

Saving output

In the *Saving output* panel, the user has to specify the directory in which the different output files have to be stored, a name to label the output files, and the format of the output files (.html, .txt, and/or .mat). The way in which the output is reported will depend on whether the user has chosen to rotate the scores (or loadings, in the case of row-wise linked data blocks) toward *A specific target matrix* or toward *All possible target matrices*:

1. ***A specific target matrix*** The output file is labeled with the name that was specified in the *Saving output* panel. The output begins by summarizing the input settings. Then the results of the analysis are shown, including the deviation of the observed sum of squared component scores from the ideal, the number of iterations used, a warning in the case that the maximal number of iterations was reached, the block-specific sums of squared scores/loadings per component before and after rotation, the rotated scores and loadings, and the rotation matrix **B**.
2. ***All possible target matrices*** Different output files are generated, one for each solution with the lowest deviation score per number of distinctive components. For each solution, the output file is labeled with the name that was specified in the *Saving output* panel, augmented with a number equal to the number of distinctive components (e.g., Psychiatry_1.txt). In each output file, the same content is presented as was described above. Furthermore, a plot of the deviation of the observed from the ideal sum of squared component scores as a function of the number of distinctive components for each target matrix is provided (see, e.g., Fig. 3 in section "Selecting an optimal target matrix"). Both figures are saved in .png and .pdf formats. In the case that a bootstrap analysis of the target selection was chosen, a histogram (in .png and .pdf formats) and table is stored in a directory called *Target Bootstrap* (which is located in the selected output directory) that displays the frequency of bootstrap samples for which the target was selected. The histograms display only targets with nonzero frequencies, and in the case of more than ten targets with nonzero frequencies, only the ten most frequently selected targets are shown.

For both options, a simultaneous component scree plot is given, which displays the proportions of VAF in each data block by each unrotated simultaneous component (see Fig. 2 in section "Selecting the number of simultaneous components"). In the case that a bootstrap analysis of the rank selection procedure was chosen, the results of this bootstrap analysis are stored in the directory *Number of components* (located in the specified output directory) both as bar charts (in .png and .pdf formats) and as a table displaying the number of bootstrap replications that resulted in the selection

of the particular rank (with the selection resulting from a parallel analysis).

Starting the analysis and error handling

The moment that one specifies a field in the DISCO-SCA program incorrectly, immediately an error message will pop up reporting the problem with the given input, and the problem is also reported in the message box. In the case that at least one field of the GUI has not been filled out properly, the *NO GO DISCO* button remains red and an overview of the inputs that are not yet specified or that are not specified properly is presented in the message box. Clicking on the *NO GO DISCO* button will report, in a pop-up message, the field or fields that still need to be specified. After correctly specifying all fields in all panels, the *GO DISCO* button turns green. This indicates that the DISCO-SCA program is ready to analyze your data. After clicking the *GO DISCO* button, the DISCO-SCA program will read the data, start the analysis, and report its progress in the message box. In the case that an error occurs while reading the data and labels, the analysis is cancelled, and an error message specifying the problem pops up. During the analysis, a *Cancel Analysis* button, which is located under the message box, will become visible. Pressing this button will cancel the analysis; in that case, no output will be saved. When the analysis has finished, a message "The analysis has been finished" will pop up. After clicking the *OK* button, the results can be consulted in the specified output directory.

Conclusion

An important goal in the analysis of linked data is to disentangle the mechanisms underlying all of the data blocks under study (i.e., common mechanisms) and the mechanisms underlying one or a few data blocks only (i.e., distinctive mechanisms). Simultaneous component analysis with rotation to DIStinctive and COMmon (DISCO-SCA) components has been proposed as a method to find such mechanisms. In this article, we have illustrated by means of a psychiatric diagnosis data set the different steps of a DISCO-SCA analysis. Furthermore, we have presented a GUI to perform a DISCO-SCA analysis. Besides being very user-friendly, this GUI also facilitates the choice of model parameters, such as the number of mechanisms and their status as being common or distinctive. Furthermore, just by a click, the data can be preprocessed in various ways. A standalone and a MATLAB version of the GUI are freely available. In this way, substantive researchers are offered an easy-to-use, all-in-one tool to support quests for common and distinctive mechanisms underlying linked data.

Author Note This work was supported by Belgian Federal Science Policy (IAP P7/06). We thank three anonymous reviewers for their constructive comments.

Appendix: Estimation of the rotation matrix

To minimize the rotation criterion in Eq. 5 for a given target, we rely on a numerical procedure called iterative majorization (see, e.g., de Leeuw, 1994; Heiser, 1995; Kiers, 1997; Lange, Hunter, & Yang, 2000; Ortega & Rheinboldt, 1970). The general set-up is as follows:

- *Step 1*: Initialize the rotation matrix $\mathbf{B}^0$, subject to $\mathbf{B}^0\mathbf{B}^{0\prime}=\mathbf{I}=\mathbf{B}^{0\prime}\mathbf{B}^0$. Define $h(\mathbf{B}^0)$ as $\|\mathbf{W} \circ (\mathbf{T}^{\text{target}}{}_{\text{conc}} - \mathbf{T}_{\text{conc}}\mathbf{B}^0)\|^2$. Initialize the iteration index: $l=1$.
- *Step 2*: Compute $\mathbf{B}^l=\mathbf{V}\mathbf{U}'$, with $\mathbf{V}$ and $\mathbf{U}$ obtained from the singular value decomposition of $\mathbf{Y}'\mathbf{T}_{\text{conc}}=\mathbf{U}\mathbf{S}\mathbf{V}'$, with $\mathbf{Y}=\mathbf{T}_{\text{conc}}\mathbf{B}^{l-1}+\mathbf{W} \circ (\mathbf{T}^{\text{target}}{}_{\text{conc}} - \mathbf{T}_{\text{conc}}\mathbf{B}^{l-1})$ (see Kiers, 1997).
- *Step 3*: Compute $h(\mathbf{B}^l)=\|\mathbf{W} \circ (\mathbf{T}_{\text{conc}}\mathbf{B}^l - \mathbf{T}^{\text{target}}{}_{\text{conc}})\|^2$. If $h(\mathbf{B}^{l-1}) - h(\mathbf{B}^l) > \varepsilon$ (with ε being a predefined small positive value; e.g., $\varepsilon = 10^{-8}$) and the maximal number of iterations is not attained, set $l=l+1$ and return to Step 2, or else consider the algorithm as having converged.

This algorithm is closely related to the gradient projection technique of Bernaards and Jennrich (2005) and converges to a stable point. To account for the fact that the algorithm may end in a local minimum, a multistart procedure may be used. Such a procedure consists of running the algorithm with several initial matrices $\mathbf{B}^0$ and retaining the solution with the lowest value of Eq. 5.

References

Alter, O., Brown, P. O., & Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences, 100,* 3351–3356. doi:10.1073/pnas.0530258100

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65,* 676–696. doi:10.1177/0013164404272507

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics, 17,* 16–33.

Browne, M. W. (1972). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25,* 115–120.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27,* 509–540. doi:10.1207/s15327906mbr2704_2

de Leeuw, J. (1994). Block relaxation algorithms in statistics. In H. Bock, W. Lenski, & M. Richter (Eds.), *Information systems and data analysis* (pp. 308–325). Berlin, Germany: Springer.

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychological Methods, 17,* 100–119. doi:10.1037/a0025385

De Roover, K., Timmerman, M. E., Mesquita, B., & Ceulemans, E. (2013). Common and cluster-specific simultaneous component analysis. *PLoS One, 8,* e62280. doi:10.1371/journal.pone.0062280

Fabrigar, L. R., Wegener, D., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299. doi:10.1037/1082-989X.4.3.272

Heiser, W. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In W. Krzanowski (Ed.), *Recent advances in descriptive multivariate analysis* (pp. 157–189). Oxford, UK: Oxford University Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185. doi:10.1007/BF02289447

Kaiser, H. F. (1958). The VARIMAX criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200. doi:10.1007/BF02289233

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika, 62,* 251–266. doi:10.1007/BF02295279

Kiers, H. A. L., & ten Berge, J. M. F. (1989). Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations. *Psychometrika, 54,* 467–473. doi:10.1007/BF02294629

Lange, K., Hunter, D. R., & Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics, 9,* 1–20. doi:10.2307/1390605

Lofstedt, T., & Trygg, J. (2011). Onplsa novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics, 25,* 441–455. doi:10.1002/cem.1388 10.1002/cem.1388

Meinshausen, N., & Buhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B, 72,* 417–473. doi:10.1111/j.1467-9868.2010.00740.x

Mezzich, J. E., & Solomon, H. (1980). *Quantitative studies in social relations*. London, UK: Academic Press.

Nishimura, Y., Martin, C. L., Vazquez-Lopez, A., Spence, S. J., Alvarez-Retuerto, A. I., Sigman, M., & Geschwind, D. H. (2007). Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human Molecular Genetics, 16,* 1682–1698. doi:10.1093/hmg/ddm116

Ortega, J. M., & Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. New York, NY: Academic Press.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis, 49,* 974–997. doi:10.1016/j.csda.2004.06.015

Rossier, J., de Stadelhofen, F. M., & Berthoud, S. (2004). The hierarchical structures of the NEO PI-R and the 16 PF 5. *European Journal of Psychological Assessment, 20,* 27–38. doi:10.1027/1015-5759.20.1.27

Saunders, D. R. (1962). Trans-varimax: Some properties of the RATIOMAX and EQUAMAX criteria for blind orthogonal rotation. *American Psychologist, 17,* 395–396.

Schouteden, M., Van Deun, K., Pattyn, S., & Van Mechelen, I. (2013). SCA and rotation to distinguish common and distinctive information in linked data. *Behavior Research Methods, 45*, 822–833. doi:10.3758/s13428-012-0295-9

Tauler, R., Smilde, A., & Kowalski, B. (1995). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics, 9,* 31–58.

ten Berge, J. M. F., Kiers, H. A. L., & Van der Stel, V. (1992). Simultaneous components analysis. *Statistica Applicata, 4,* 377–392.

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology, 59,* 301–320. doi:10.1348/000711005X67599

Timmerman, M. E., & Kiers, H. A. L. (2003). Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika, 68,* 105–121. doi:10.1007/BF02296656

van den Berg, R., Van Mechelen, I., Wilderjans, T., Van Deun, K., Kiers, H., & Smilde, A. (2009). Integrating functional genomics data using maximum likelihood based simultaneous component analysis. *BMC Bioinformatics, 10,* 340. doi:10.1186/1471-2105-10-340

Van Deun, K., Smilde, A. K., van der Werf, M. J., Kiers, H. A. L., & Van Mechelen, I. (2009). A structured overview of simultaneous component based data integration. *BMC Bioinformatics, 10,* 246–261. doi:10.1186/1471-2105-10-246

Van Deun, K., Van Mechelen, I., Thorrez, L., Schouteden, M., De Moor, M., van der Werf, M. J., & Kiers, H. A. L. (2012). DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes. *PLoS One, 7,* e37840. doi:10.1371/journal.pone.0037840

Van Deun, K., Wilderjans, T. F., van den Berg, R. A., Antoniadis, A., & Van Mechelen, I. (2011). A flexible framework for sparse simultaneous component based data integration. *BMC Bioinformatics, 12,* 448. doi:10.1186/1471-2105-12-448

Wilderjans, T., Ceulemans, E., & Van Mechelen, I. (2009a). Simultaneous analysis of coupled data blocks differing in size: A comparison of two weighting schemes. *Computational Statistics and Data Analysis, 53,* 1086–1098. doi:10.1016/j.csda.2008.09.031

Wilderjans, T., Ceulemans, E., Van Mechelen, I., & van den Berg, R. (2009b). Simultaneous analysis of coupled data matrices subject to different amounts of noise. *British Journal of Mathematical and Statistical Psychology, 64,* 277–290. doi:10.1348/000711010X513263